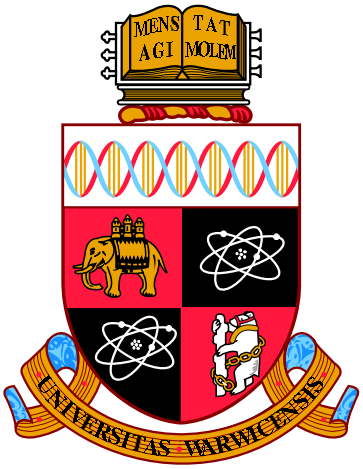


Taming Web Data: Exploiting Linked Data for Integrating Medical Educational Content

by

Reem Ali Qadan Al Fayez

A thesis submitted in partial fulfilment of the requirements for the

degree of

Doctor of Philosophy

Supervisor: Dr. Mike Joy

Department of Computer Science

February 2016

This thesis is dedicated to my parents.
For their endless love, support, and encouragement

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I would like to express my deepest gratitude to God who gave me the strength and courage needed to complete my studies. The amount of work and effort put into this thesis is tremendous. My heartfelt appreciation goes to my supervisor Dr. Mike Joy as without his constant support and encouragement this work would not have been possible. His guidance through the years of my study has gave me the confidence to pursue in the same momentum I started four years ago.

My family, Mom, Dad, my brothers, and sisters, all the thank you words will never give your right. Not only you gave me all the support and encouragement I need to tolerate the burden of the PhD, but you also had to put up with the fact that I missed so many family events, happy or sad, and I could not be part of them. I love you is more expressive than thank you, as you drowned me with all the love and support any person can ever have. The journey have started four years ago, and it has been the greatest experience I ever had. I owe a very important debt to all the friendships I made during this journey. My friends you made this journey an unforgeable one. Thank you for each and every one of you. How far or close you were to me, you all had an impact on my journey. You were standing by my side through all the ups and downs I had. I was always certain that I had a solid back to lean on and a circle of support that I can turn to whenever I want. My friends, whether I knew you for a month, a year, or four years, make sure that you all had shaped me into the person I am today, and you taught me things about life and about myself. I will always look back on this journey with a wide smile on my face because of each and everyone of you, my supervisor, my family, my colleagues, and my friends.

Publications

The publications written during the PhD research are listed below and their connection to the research presented in this thesis is explained.

1. At an early stage of this research, the main research area explored was the adaptive delivery of educational content via social network. Therefore, we researched how can the User Generated Content (UGC) provided by users of micro-blogging services, in particular Twitter, be used for disambiguating terms and inferring possible taxonomies for terms that will aid in building adaptive services for matching and delivering of educational content to users. The work was written as a paper published in **the INTECH** conference 2013.

    - Al Fayez, R.Q. and Joy, M., 2013, August. Inferring dynamic taxonomies for terms based on UGC. In Innovative Computing Technology (INTECH), 2013 Third International Conference on (pp. 545-550). IEEE. Vancouver.

2. The research shifted towards integrating educational objects from Web 2.0 websites and online libraries into one linked dataset. Hence, we proposed the LEMO metadata schema that accommodates describing different types of educational medical objects and enrich their metadata with biomedical ontology concepts that will be used for linking the distributed objects collected. The LEMO metadata schema is presented in chapter 4 and was published in **the Web Information System Engineering conference (WISE)** 2014.

    - Al Fayez, R.Q. and Joy, M., 2014. A framework for linking educational medical objects: connecting web2. 0 and traditional education. In Web Information Systems EngineeringWISE 2014 (pp. 158-167). Springer International Publishing

3. The researched continued to test and developed a linked dataset of educational medical objects described using the LEMO metadata schema. Hence,

we proposed the LEMO system that enables harvesting, mapping, and enriching heterogeneous metadata of object collected from the web into one linked dataset. The LEMO system is presented in chapter 5 and experiments of using this system and evaluating its results are presented in chapters 6, chapter 7, and chapter 8. Part of this work was published in the **British International Conference On Databases (BICOD)** 2015.

- Al Fayez, R.Q. and Joy, M., 2015. Applying NoSQL Databases for Integrating Web Educational Stores-An Ontology-Based Approach. In Data Science (pp. 29-40). Springer International Publishing.

4. The BICOD paper has been selected as one of the best papers from BICOD 2015 and we were invited for submitting an extended version for **the Computer Journal** special issue featuring the best papers from BICOD 2015. The extended journal paper titled " Using Linked Data for Integrating Educational Medical Web Databases based on BioMedical Ontologies" has been written to summarise the overall research presented in this thesis. The extended journal paper has been examined by referees and by the guest editors and first decision of acceptance has been given with referees comments that were addressed and submitted in 14th of February 2016.

Abstract

Open data are playing a vital role in different communities, including governments, businesses, and education. This revolution has had a high impact on the education field. Recently, new practices are being adopted for publishing and connecting data on the web, known as "Linked Data", and these are used to expose and connect data which were not previously linked. In the context of education, applying Linked Data practices to the growing amount of open data used for learning is potentially highly beneficial. The work presented in this thesis tackles the challenges of data acquisition and integration from distributed web data sources into one linked dataset. The application domain of this thesis is medical education, and the focus is on bridging the gap between articles published in online educational libraries and content published on Web 2.0 platforms that can be used for education. The integration of a collection of heterogeneous resources is to create links between data collected from distributed web data sources. To address these challenges, a system is proposed that exploits the Linked Data for building a metadata schema in XML/RDF format for describing resources and enriching it with external dataset that adds semantic to its metadata. The proposed system collects resources from distributed data sources on the web and enriches their metadata with concepts from biomedical ontologies, such as SNOMED CT, that enable its linking. The final result of building this system is a linked dataset of more than 10,000 resources collected from PubMed Library, *YouTube* channels, and Blogging platforms. The effectiveness of the system proposed is evaluated by validating the content of the linked dataset when accessed and retrieved. Ontology-based techniques have been developed for browsing and querying the linked dataset resulting from the system proposed. Experiments have been conducted to simulate users' access to the linked dataset and validate its content. The results were promising and have shown the effectiveness of using SNOMED CT for integrating distributed resources from diverse web data sources.

Chapter 1

Introduction

The World Wide Web (WWW) has become the primary source of information considering the widespread use of the internet. Initiatives such as open source (in development), open access (in research), open content (in publications), and the most recent open data movement, all have similar goals. The overall intention of the "open data" movement, which gained its popularity with the launch of open-data government, is to make some data available for everyone to use and republish freely. The Open Definition[1] has specified the meaning of the term open as "Anyone can freely access, use, modify, and share for any purpose". Hence, the qualifier "Open" has been associated with different communities where its proponents believe that the elimination of barriers can encourage its opportunities and participation. Open education is used as a collective term to accommodate all the initiatives that promote free education, such as Open Educational Resources (OER) [Iiyoshi and Kumar, 2008], and such initiatives are changing education. For example, books used in schools and universities are becoming e-books, articles are being published as open access, while some educational institutes videotape lectures and publish them on *YouTube*. These are few examples of wider educational content made available freely on the web. Therefore, it became the norm for both students and educators

[1] http://opendefinition.org/

to depend on the web for acquiring knowledge. Furthermore, research had showed that utilising educational content published using Web 2.0 technologies supports the learning process [Newland and Byles, 2014]. Examples of such education content are videos, blogs, wikis, or pictures. This trend applies to all fields of education whether it is humanities, scientific, or medical. However, the large volumes of open data made available on the web put a higher burden on the user when searching. The process of finding information published in trustworthy web sources requires both effort and time from users. Learners tend to search for different types of materials such as videos, pictures, and blog articles to aid them in understanding concepts they are studying in books and during lectures. Existing search engines can be frustrating to use for this purpose. Some are not designed for educational purposes such as *Google* while others might limit the search to a collection of materials they are hosting as in universities' libraries. Considering all the above issues, learners of any field of study face the same problem of finding the information they need on the web to help them study. Therefore, this thesis addressed the search problem in the field of medical education since various technologies have emerged for enhancing the learning and teaching experience, and have been incorporated for developing medical e-curricula [Fleiszer et al., 2004]. Besides, research in the field of medical education provided insights into the potential impact of Web 2.0 technologies on enhancing teaching and learning [Popoiu et al., 2012].

The process of searching any web data source is thus twofold. Firstly, the search process is made easier for users if the published content is described using representative metadata and thus provides what it needs for matching any search query. Secondly, the user's search query must represent the information the user seeks, and that affects the correctness of the search results. The data fields describing any piece of data published on the web is named "Metadata". Different metadata models have been implemented by organizations such as IEEE to accommodate the requirements of publishing their content [Sampson, 2004]. By having too many

metadata models proposed in the field of education, it became evident that no ideal standard accommodates the needs of all publishing organizations. It is also the case in the medical education since several libraries hosting medical education content had proposed and used data models and vocabularies for organizing their content. Research had been investigating the organization of educational content in online libraries, but fewer efforts focused on integrating their content. With the emerging use of Web 2.0 content in learning, research has to focus not only on integrating web data sources of traditional libraries, but also on integrating web data sources from hosting different types of content including Web 2.0 content.

Now that the way of publishing data on the web is changing, it is possible for datasets collected from distributed web sources to be exposed on the web and linked together. The new vision of the web is growing into building what is called the "Web of Data" that facilitate Web-scale data integration using new standards for publishing data on the web. This new web aims to provide machine-readable data that can be semantically enriched to enable linking data published in distributed datasets [Heath, 2008]. Semantic web community has been developing techniques and methods for making the "Web of Data" feasible. They developed new practice named the "Linked Data" practice that adopts existing web technologies to expose data that is already published on the web. Linked Data can be defined as data published on the web using URIs and Resource Description Framework (RDF) that make it machine-readable, semantically defined, easily linked to external datasets, and can be linked to from external datasets [Bizer, Heath and Berners-Lee, 2009]. Therefore, Linked Data practice is used to expose and share data on the web where it lowers the barriers to building links between the data. The adoption of Linked Data practice is increasingly turning the web into a global data space [Heath and Bizer, 2011]. In education, the use of Linked Data is becoming popular [Vega-Gorgojo et al., 2015]. Exploiting the features of linked data for publishing open educational content enables the integration of web data sources hosting content of heterogeneous

nature on a full web-scale.

1.1 Problem Definition

The work presented in this thesis illustrates how to tackle the challenges of data acquisition and integration into appropriate presentation and organization with web data in the context of medical education. In particular, this work focuses on bridging the gap between the content of online educational libraries and Web 2.0 that are both used in learning. In 2002, IEEE launched the IEEE LOM standard that defined a Learning Object as "any entity, digital or non-digital, that may be used for learning, education or training" [Learning Technology Standards Committee, 2002]. The two terms "Learning Object" and "Educational Object" have been used interchangeably in the same meaning [Friesen, 2001]. Therefore, throughout this thesis, the term Educational Medical Objects (EMOs) is used to refer to educational content collected from different web data sources and the collection of related EMOs after integration is named the Linked Educational Medical Objects (LEMO) dataset.

The integration of EMOs is to create links between heterogeneous data collected from distributed web data sources. The integration process is performed by exploiting Linked Data practices for exposing and linking the EMOs. Before the integration process, each EMO published in a web data source is represented using structured metadata elements that provide details about its title, authors, description, data of publishing, and other attributes supported by that metadata. As the structure of metadata elements differs from one web data source to another, the process of integrating EMOs is hard to achieve.

Another issue discussed in this work is accessing and retrieving data from the LEMO dataset resulted from aggregating distributed web data sources. Since the EMOs are aggregated from various web data sources and the aggregation process is performed without any human involvement, two issues arise that affect accessing

and retrieving data from the LEMO dataset. First, the LEMO dataset can have EMOs explaining a broad range of topics in the field of medical education that can not be predefined. The second issue is that some EMOs aggregated might have poorly described metadata making it hard to retrieve when searching which might be the case for EMOs aggregated from Web 2.0 data sources where the metadata is user-generated content.

1.1.1 Research Motivation

In this era of open data, experts are sharing their knowledge via blogs, academic institutions are providing full courses for free, and journals are being published with open access to its article. Thus, public web resources that can be used for learning are increasingly plentiful. Such practices can significantly impact the learning and teaching habits.

For example, a high school student has to prepare a presentation about "Breast Cancer" for Biology class. The presentation must include demonstrating pictures, medical cases, in addition to the main text explaining the topic presented with proper referencing. If the student wants to complete this task successfully, he/she has to do enough research to understand the topic and find the right resources to include in the presentation. High school students might not have the privilege to access medical journals, so he/she has to search for open web data sources to complete this task. In this fast pace of life, students prefer to read small articles or watch videos in order to gain an overview of the topic. For this task, the student might start preparing for the presentation by searching *YouTube*[2] or *SlideShare*[3]. Then, he/she would search the topic in *Google*[4] and read news articles, experts' blogs, and possibly patients' blogs. Having had an overview of the topic, the student can start searching for books and peer-reviewed articles to complete the presentation.

[2]http://www.youtube.com
[3]http://www.slideshare.net
[4]http://www.google.coom

At the end of the search process, the student should have collected appropriate materials to include in the presentation and accomplish the task successfully.

From this example, it can be noticed that the search process consumes tremendous amount of time and effort from learners to acquire the knowledge he/she desires about a topic. Besides, the increased use of public web resources in learning and teaching motivates this research and encourage finding a practical solution that can be applied. The work presented in this thesis is applied to the medical education field as a proof of concept that can be extended to include different fields if it is proved to be valid.

1.1.2 Research Challenges

The challenges in integrating the data from heterogeneous web data sources include the diversity of metadata schemas involved, the quality of the metadata descriptions provided, and the ambiguity of the topics of the data aggregated from these web data sources. These are the limitations of this research with further details uncovered in the upcoming chapters.

Web 2.0 hosting websites such as *YouTube* have metadata elements that are used to describe its videos. Also, blogs are published in different blogging platforms that might vary in their metadata elements describing their blogs. Comparing such metadata with the metadata schemas, used to organize the books and articles published in online libraries, assures the heterogeneity of possible metadata schema involved in this research. Another challenge is the quality of the metadata describing the basic information about the EMOs. The incompleteness of some metadata elements affects its quality such as having objects with no title attributes filled. Also, metadata that does not represent the actual content of the EMO described affects the process of integrating or retrieving that EMO. For example, publishing a video on *YouTube* and not including a proper description of its content in its metadata limits the search for that video. Finally, aggregating EMOs from distributed

web data sources without having restrictions on the topics of these EMOs presents a challenge on organizing and linking various EMOs explaining different topics in the domain of medical education. The following list summarizes the main challenges of this research and possible techniques applied to tackle these challenges.

- Metadata schemas: The diversity of metadata schemas applied for describing the content published in web data sources must be taken into account. The solution proposed in this research has to consider mapping all these metadata elements into one unified structure.
- Incompleteness of metadata fields: Techniques must be developed as part of the solution proposed for enriching the description of EMOs by adding semantics to it.
- Unsupervised aggregation of content: Automatic organization of the data aggregated has to be part of the solution proposed. Techniques must be developed to handle the diversity of topics that might be covered by EMOs aggregated from the web data sources such as automatic tagging and categorizing of the EMOs.

1.2 Research Questions and Objectives

The research focuses on proposing a solution for solving the problem identified in the previous section. The research questions and objectives are detailed in this section. Further details are presented in the following chapters for the work conducted to achieve these objectives and answer these questions.

1.2.1 Research Questions

The work presented in this thesis proposes techniques and methods to answer this research question. The main research question is broad and can be further subdivided into three related questions. The main research question to answer is:

R0: How can Linked Data be used to support the acquisition and integration of EMOs from distributed web data sources? The EMOs represent any piece of information that can be used for learning in the field of medical education such as articles, videos, and blogs.

The research conducted to answer this question, raised multiple related questions mapped to the research objectives, that support answering these questions, that are detailed in the next section. The following questions are the main focus of this thesis.

R1: What are the current metadata schemas used when publishing medical educational content on the web and what are the essential elements from the user perspective when searching for such content? (O1, O2).

R2: How can Linked Data practice be used to design and implement a metadata schema that accommodates various types of EMOs and enables its exposing and linking with the aid of external datasets such as biomedical ontologies? (O3, O4).

R3: What are the techniques used for harvesting, mapping, and organising EMOs from distributed web data sources into one linked dataset? (O5, O6, O7).

R4: How can the Linked Data practices be utilised in the process of accessing and querying the dataset of integrated EMOs called the LEMO dataset? Moreover, how can the linkages between content retrieved from the LEMO dataset be evaluated? (O9, O10).

1.2.2 Research Objectives

The objectives of this work are to aggregate and integrate Educational Medical Objects (EMOs) from distributed web data sources into a linked dataset that can be accessed and queried. The aspects addressed by this work are divided into two

problems. Firstly, the problem of collecting EMOs with heterogeneous metadata formats and organising them into one dataset. Secondly, the issue of building a linked dataset where connections are created between possibly related EMOs. In this thesis, the first problem is addressed by proposing a metadata schema that accommodates various metadata schemas named the Linked Educational Medical Objects (LEMO) Application Profile. While the second problem of building a linked dataset of EMOs is addressed by developing a system that exploits the Linked Data practice for exposing and connecting the EMOs using the web. The system is named the LEMO system. The system experiments with different biomedical ontologies and is evaluated using real web data sources content collected from *YouTube*, *Blogging platforms*, and *PubMed Library*.

The detailed research objectives are listed as follows. The chapters of this thesis aim to achieve these objectives.

O1: Identify existing metadata schemas that are already used for describing EMOs in specialized medical educational libraries.

O2: Identify the search practices and challenges faced by students and educators in the medical education field when searching for educational content on the web including Web 2.0 sites and online academic libraries.

O3: Conduct a comparative analysis of the existing metadata schemas to identify the common characteristics for describing EMOs.

O4: Design the proposed LEMO metadata schema by introducing new features to enrich the description of EMOs and enable its integration into one linked dataset. Then, implement the LEMO metadata schema in Linked Data format, and validate that metadata schema by conducting experiments for describing real EMOs of different types that are collected from diverse web data sources.

O5: Establish the LEMO system framework for harvesting, mapping, and interlinking EMOs using Linked Data techniques.

O6: Identify and develop possible techniques and protocols used for harvesting and mapping EMOs from web data sources and describe it using the LEMO metadata schema.

O7: Investigate possible biomedical ontologies that can be used to expose and interlink the EMOs, and develop tools for exploiting the biomedical ontologies in enriching the EMOs metadata with semantics to enable its integration into one linked dataset.

O8: Describe the RDF store that is managed by the LEMO system for organising the EMOs metadata represented in the LEMO metadata schema.

O9: Develop ontology-based method for browsing the LEMO dataset resulted from the LEMO system and evaluate the similarity between the results retrieved while browsing.

O10: Develop an ontology-based query searching algorithm for testing and comparing of query searching results between ontology-based and text-based searching methods in the LEMO dataset.

1.3 Research Design Methodology

As stated in the problem definition (section 1.1), this research focuses on exploiting Linked Data practice to tackle the challenges of data acquisition and integration from web data sources. The methodology followed to conduct this research is split into three phases. At an early stage of this work, an exploratory research is started that included studying a sample domain representing the medical education community and conducting a background research that covered the needed information to propose a solution for this problem. The recommendations and discoveries resulted

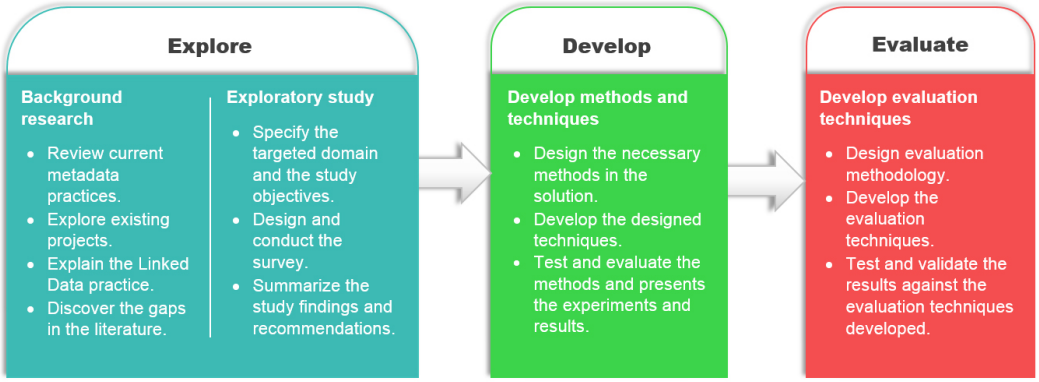


Figure 1.1: The research design methodology

from the exploratory research are the input for the next phase of designing and developing the solution proposed in this thesis. The final phase is testing and evaluating the results of implementing the solution proposed. The research methodology followed in this thesis is illustrated in Figure 1.1.

1. Exploratory research:

   Since the field of Linked Data is considered a modern research area and its applications in education is still mature, a comprehensive background research concerning the field of linked data and its application is conducted in addition to further research areas. Also, an exploratory study was undertaken to explore the requirements and problems identified from the perspective of the medical education community.

   - Background research: The work presented in this thesis relates different research areas together. Since the problem addressed in this research is concerned with the medical education field, the literature review starts with investigating current practices in dedicated medical educational libraries. It also explores existing projects being developed for organizing medical educational materials. Furthermore, the background research investigates the current techniques and methods for publishing content on

the web since new practices such as Linked Data are adopted in this work. Also, the background research explores current projects and applications conducted that exploit Linked Data on issues concerning education in general and medical education in specific. The gaps discovered in the background research and the necessary background knowledge required to understand this work are presented in Chapter 2.

- Exploratory study: The problem identified in this research is applied to the field of medical education. An exploratory study is conducted with participants from the medical education community. A sample of students and educators had represented the medical education community and identified the problems and frustrations from their perspective when searching the web. The results of this study have been presented in Chapter 3 and further recommendations are suggested to support the design of the proposed solution.

2. Development:

   The gaps discovered in the background research and the recommendations resulted from the exploratory study are the input for this phase. Both are used to identify the functional requirements for designing and developing the LEMO metadata schema and the LEMO system. The proposed solution is designed and developed in this phase. The detailed process of proposing and implementing the LEMO metadata schema is explained in Chapter 4. The LEMO system is then implemented to exploit the features proposed in the LEMO metadata schema and build a linked dataset of EMOs. The design and implementation of the LEMO system are presented in Chapter 5.

3. Evaluation:

   The LEMO system is tested and evaluated against real data harvested from web data sources. The final phase of this research displays the results of

running the LEMO system to build a linked dataset of EMOs. This phase presents the final linked dataset named the LEMO dataset in Chapter 6. The evaluation of the LEMO dataset is conducted via developed techniques for accessing and retrieving its objects. Browsing techniques are developed and tested in Chapter 7 to explore the LEMO dataset content and evaluate its linkages. Also, the LEMO dataset is tested via query searching endpoint developed and evaluated in Chapter 8.

1.4 Research Contributions

The main contributions of the work presented in this research are summarized in the following subsections. Each subsection gives an overview of the contribution and the chapter that details it.

1.4.1 The LEMO Metadata Schema

In Chapter 4, a metadata schema is proposed for describing the different types of EMOs collected from various web data sources. The metadata is named Linked Educational Medical Object (LEMO) because it is implemented using Linked Data practice. The proposed LEMO metadata aims to bridge the gap between Web2.0 data sources and online educational libraries by accommodating the description of EMOs collected from both data sources. Therefore, the first step before proposing the LEMO metadata was analysing the metadata schemas applied for describing all types of EMOs involved in this research. Then, the LEMO data model is proposed by extending Dublin Core Metadata Initiative (DCMI) schema [Powell et al., 2007] instead of building a new one from scratch. The design process of the LEMO metadata explains in detail the functional requirements elicitation process including a justification for deciding to extend the DCMI schema. The proposed LEMO metadata introduces new elements that enable enriching the description of EMOs

with semantics using external biomedical ontologies. Linked Data practice is applied to implement the LEMO metadata and enable building links between its EMOs. It is developed in RDF/XML format with further techniques developed for mapping the heterogeneous metadata schemas collected into the LEMO metadata. To test the practicality of the LEMO metadata, several experiments were conducted with real data that represent EMOs collected from various web data sources.

1.4.2 The LEMO System

Prompted by the newly introduced features proposed in the LEMO metadata schema explained in Chapter 4, the LEMO system has been developed. The design and implementation processes of the LEMO system are detailed in Chapter 5, where the architectural decisions, the system architecture, and the system implementation are explained. The LEMO system provides components for harvesting and mapping EMOs from distributed web data sources into the LEMO metadata schema forming the LEMO dataset. Moreover, it developed methods for enriching the EMOs by annotating some elements of the LEMO metadata with concepts from external biomedical ontologies. Using the enriched LEMO dataset, the system developed techniques for categorizing the EMOs in the dataset by proposing algorithms that utilize the hierarchical relations between concepts in the biomedical ontologies. Experiments were conducted to test and compare the use of two well-known biomedical ontologies for enriching the LEMO dataset. The results of these experiments illustrate how links can be built between the EMOs based on the semantics added to the LEMO dataset when enriched.

1.4.3 The LEMO dataset

The outcome of running the LEMO system that aggregate EMOs from distributed web data sources and integrate them based on the LEMO metadata, is described as one linked data set of EMOs called the LEMO dataset. A large-scale experiment

was conducted to build the LEMO dataset that can be accessed and evaluated later in the research. The LEMO dataset is presented in Chapter 6. It included different type of EMOs such as videos harvested from *YouTube*, blogs harvested from *Journal Blogs*, in addition to EMOs that represent articles harvested from the *PubMed library*. The final LEMO dataset is stored in an RDF store that is presented in detail with further explanation of the added enrichments and generated linkages between its components.

1.4.4 Ontology-based Retrieval

The work presented in this thesis proposes ontology-based information retrieval paradigms for evaluating the resulted LEMO dataset presented in chapter 6. A comprehensive evaluation of the linked dataset is necessary to validate the techniques and methods developed in the previous chapters. In chapter 7, browsing navigation method is presented in detail. While, chapter 8 presents another information retrieval paradigm that is query searching. Both retrieval methods utilise the ontology used for enriching the metadata elements of the LEMO dataset. Evaluation techniques were developed to simulate users' access to the LEMO dataset. Browsing the LEMO dataset explores its content and provides an overview of the linkages created between its EMOs. The evaluation of the browsing techniques developed included clustering experiments of the browsing results that indicate the efficiency of the LEMO dataset organisation. As for query searching, an ontology-based algorithm is proposed and tested using random queries that simulate users' query actions. Compared with simple text-based retrieval, the results of the ontology-based query searching indicate an improved discovery and retrieval of EMOs from the LEMO dataset.

1.5 Thesis Outline

This chapter introduced the research problem, the challenges, the research objectives and the research questions. In total, this thesis consists of nine chapters (including the current chapter), organized as follows.

Exploratory research conducted is reported in two chapters. **Chapter 2** elaborates on the background research needed to understand the work presented in this thesis, while **Chapter 3** details the results of an exploratory study performed with participants from the field of medical education. Based on the gaps discovered in the background research and the recommendations from the exploratory study in the domain studied, the development of the solution proposed in this thesis is presented in the following two chapters. The LEMO metadata schema is designed and developed in **Chapter 4**, followed by the development of the LEMO system in **Chapter 5** that is designed to exploit the features of the LEMO metadata schema for solving the problem discussed in this research. **Chapter 6**, **Chapter 7**, and **Chapter 8** detail the results and evaluation of the work conducted in this thesis. **Chapter 6** presents the final dataset and its components with further explanation of its organization supported by a detailed example of one EMO and how it is stored in the LEMO dataset. **Chapter 7** develops and evaluate the browsing technique of the LEMO dataset, while **Chapter 8** explains and validate the technique developed for querying the LEMO dataset. Finally, **Chapter 9** concludes the thesis and provides information about potential future work.

Chapter 2

Background and Related Work

2.1 Introduction

This chapter provides the foundation needed for better understanding of the work presented in this thesis. It consists of background information about two topics that are relevant to understanding this research. Furthermore, it provides a discussion of the related work that has been accomplished so far in the field of medical education that is the domain of application in this research. This part of the chapter addresses the research objective **O1**: "Identify existing metadata schemas that are already used for describing EMOs in specialized medical educational libraries". This chapter begins with detailing the need for metadata standards and the widely used standards applied in e-learning, followed by a discussion for specific metadata standards developed for organising educational content in the field of medical education. Next, a background is given about the process of publishing data on the web from the past until today. The focus of this chapter is the technique of Linked Data that is emerging to change the way the web is formed. This is followed with a discussion of the related work carried out in the literature that is concerned with applying Linked Data for organising the educational content on the web. As a result, the gaps discovered in the literature are discussed.

2.2 Metadata Standards

Traditionally, the use of metadata has been particularly important in database management systems. Nowadays, metadata has gained its importance from the essential need for data in all disciplines. Usually, metadata is structured information that is used to describe a resource to enable its discovery. Metadata facilitate the discovery and retrieval of relevant information from databases, libraries, system, and the web. The importance of metadata on the web arise from the vast heterogeneous objects that are published there, such as HTML documents, videos, pictures, just to name a few. Each type of these resources requires having descriptive information provided as metadata to improve its retrieval. On the other hand, each metadata schema proposed and used by an organisation is developed with a particular aim in mind. The general purposes of using the metadata were discussed in the literature, and the following list summarises the most common aims [de Carvalho Moura et al., 1998].

- *Bibliographic cataloguing:* one primary usage for metadata is considering them as surrogates of the real thing [Coyle, 2005]. In other words, metadata are used to organise the digital resources and index them in digital repositories according to their content. Therefore, metadata standards replaced existing cataloguing standards that were developed in the early stages of libraries automation such as the MAchine Readable Cataloguing (MARC) standards and the Anglo-American Cataloguing Rules (AACR) [Gorman, 2003]. The process of maintaining the metadata in such cataloguing standards is the job of librarians only. Therefore, using such standards in digital libraries is not efficient as it requires time and qualified staff and can be used only on a small number of resources [Gorman, 2003]. With the evolution of the internet and the increased amount of online resources, simpler metadata formats were introduced to address the issue of cataloguing resources in online libraries. Simpler metadata standards paved the way for allowing the authors of online resources to create

and maintain their metadata instead of trained professionals [El-Sherbini and Klim, 2004].

- *Reuse and interchange:* metadata are used to support the interchange and reuse of resources between different information systems [Neiswender and Montgomery, 2009]. The exchange of metadata between such systems must be performed with minimal loss of information, but having different metadata standards employed for describing resources, the interchange of information is not straightforward. Various techniques have been developed to tackle the problem of metadata interoperability such as introducing the concept of Application Profiles (APs). Instead of developing a new metadata schema, APs enables tailoring an existing metadata schema to satisfy the requirements of a particular organisation or information system [Haslhofer and Klas, 2010]. Metadata standards applied in e-learning have been among the very fast learning technology standards to mature [Nilsson et al., 2007]. Examples of such metadata standards are detailed in the following section.

- *Discovering resources on the web:* resources published on the web must be described with meaningful information to improve their retrieval. Search tools can only access metadata that are provided as surrogates of the resources' content and can not access and search the content of the resources. Metadata on the web are implemented in formats such as XML or RDF [Duval et al., 2002]. Such formats are machine-readable allowing the information systems and the search tools on the web to discover and retrieve data easily [Neiswender and Montgomery, 2011]. Web metadata that is machine-readable that is not machine-understandable unless it is semantically annotated with ontologies. Such metadata can be described in RDF formats that enable enriching the content with external data [Lassila, 1998].

The limited number of metadata standards that were chosen for discussion here does not exhaust all other metadata standards that exist, but it is intended to illustrate some of the well-known metadata standards and the purpose of that metadata.

2.2.1 IEEE LOM

The Learning Technology Standards Committee (LTSC) of the IEEE started to develop Learning Object Metadata (LOM) standard in 1997. With the support of different international participants, LOM working group succeeded to have this standard accredited by IEEE in June 2002 [IEEE LTSC, 2002] and from that point, IEEE LTSC LOM working group continued to develop and maintain this standard. Furthermore, their work extended to collaborate with other interested parties to support the interoperability of Learning Objects implementing different standards. As defined in LOM standard, an instance of LOM is designed to record the characteristics of the learning object it describes grouped into nine categories: general, life cycle, meta-metadata, educational, technical, rights, relation, annotation, and classification. The LOM standard schema is illustrated in Figure 2.1. In each of these categories, there is a set of data elements which compose as a whole a metadata instance describing a learning object. The purpose of such detailed schema, as stated by the working group of LOM, is to facilitate sharing and exchange of learning objects since the metadata has a high degree of semantic interoperability [IEEE LTSC, 2002]. The total number of elements composing LOM standard is 45 elements which are all optional to complete in this standard. They are all directly descendent from the nine parent categories and can be leaf nodes or parent nodes of further sub-elements.

Despite the fact that all the LOM elements are optional to fill when describing a learning object, there are restrictions on the values entered if filled. The LOM metadata elements have values that are governed by rules and restrictions defined in

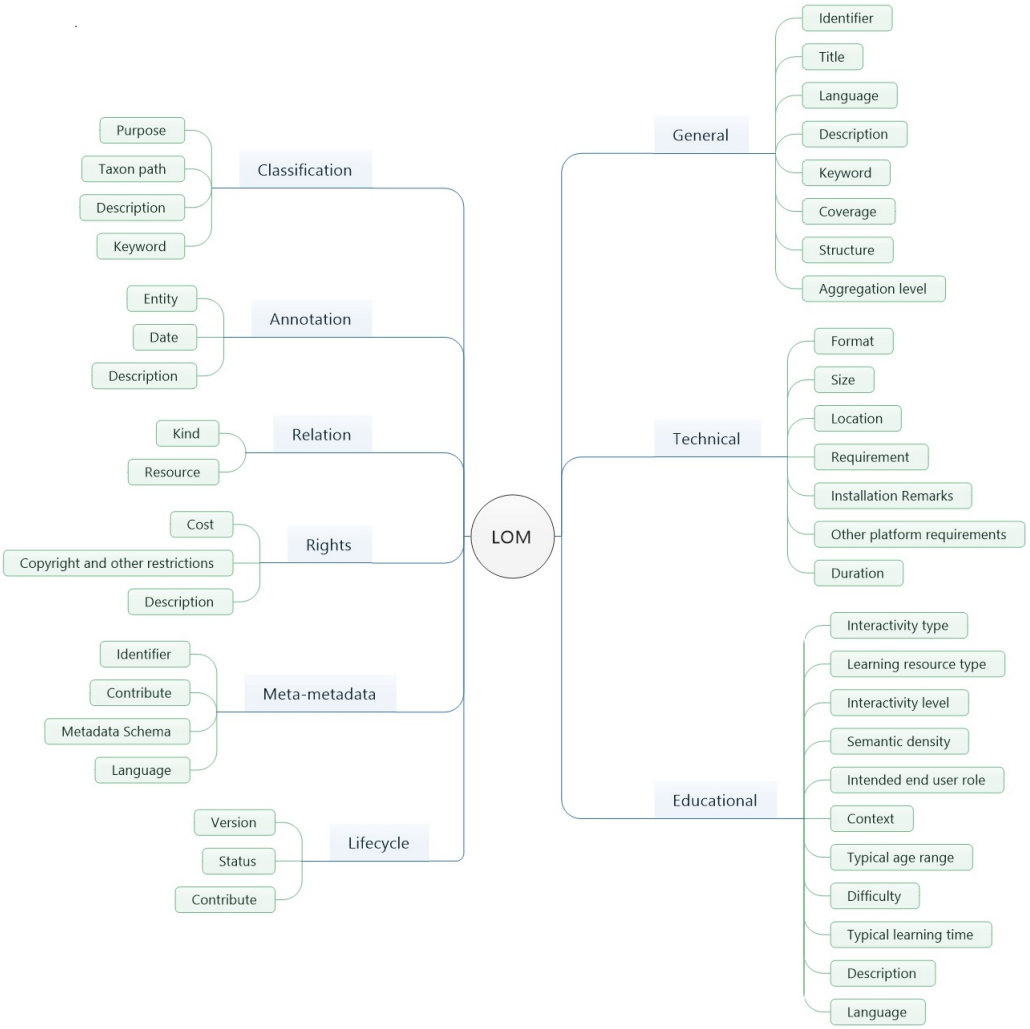


Figure 2.1: The LOM metadata schema

the standard such as the data type of its values or the size of the allowed values filling its elements. For example, a property named *smallest permitted maximum value* forces the application implementing this standard to conform the maximum size of the values entered, for a particular element, to the smallest permitted size specified by the standard. Another restriction imposed by LOM standards is the data type of the values that can be specified using *dateTime*, *duration*, and *langString* data types. Furthermore, the standard can define the language of the metadata entered

for any element using the *langString* property to store the language code. Also, the values of some metadata elements may be restricted to a controlled vocabulary, and the value of the element must be selected from that vocabulary. These are some of the rules and restrictions implemented in LOM standard that represent its complexity. Additionally, some of the elements in LOM may be repeated in different categories. For example, notice the element *keyword* in the *General* category and the *keyword* in the *Classification* category as illustrated in Figure 2.1. Hence, the same element might be used for different purposes determined by its context. Despite the complexity of this metadata standards, LOM was primarily proposed for educational purposes that aim to describe Learning Object (LOs). Therefore, the *Educational* category in the LOM standards consists of data elements that provide information about the educational use of a learning object as detailed in Figure 2.1. Other metadata standards were developed with no restrictions on the domain or the purpose for its usage such as the DCMI metadata that is widely adopted across the web.

2.2.2 DCMI

Dublin Core Metadata Initiative (DCMI) is an open public organisation that is non-profitable. It supports metadata design and implementation across a broad range of purposes. The initiative's work resulted in a simple cross-domain metadata standard known as Dublin Core Metadata Element Set (DCMES) which has been standardized as ISO standard 15836:2009 [ISO/TC, 2009] . The Dublin Core standard is used to describe a broad range of resources, where a resource is defined by DCMI as "anything which might be identified" [Kunze and Baker, 2007]. The Dublin Core (DC) metadata schema is simple and consists of the following 15 elements as illustrated in Figure 2.2. The DCMES elements are the identifier, title, description, date, creator, contributor, publisher, coverage, format, language, relation, rights, source, subject, and type. All of these elements are optional and maybe repeated

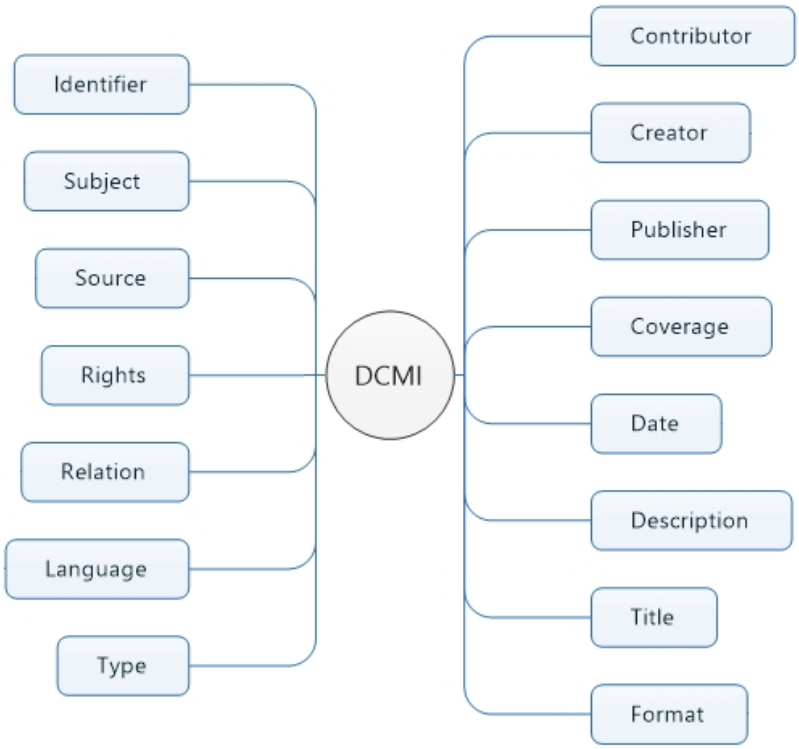


Figure 2.2: The DCMI metadata schema

if required when describing a single resource. The work on DC metadata schema started in 1998. The DCMI continued its work on developing these simple elements set to accommodate the needs of the semantic web and enable the interoperability of resources published on the web. In 2008, DCMI published new specifications for using the DCMES element set in another format that complies with the semantic web needs. The new version of the DC metadata terms was specified in RDF format and consist of 15 new properties identical to the old 15 elements in the DCMES. Hence, the interoperability of existing implementations of the DCMES will not be affected.

2.3 E-Learning Metadata Standards in Health Care

Different healthcare organisations started developing metadata standards to accommodate their requirements for publishing resources online. It is possible that existing metadata standards do not satisfy the needs and demands of a particular organisation. Thus, the concept of Application Profiles (AP) emerged in the field of metadata standardisation. Developing an Application Profile is a flexible way to adapt an existing standard and use parts of its elements to satisfy the organisational needs. Any system or organisation can implement a modified version of any metadata schema, either by adding or removing elements to the metadata schema and still guarantee the interoperability of its content [Rebai et al., 2008]. The following metadata schemas are types of application profiles developed based on existing standards for organisations in the healthcare field.

2.3.1 Healthcare LOM

HealthCare LOM metadata schema was developed by the MedBiquitous Learning Object Working Group. It is an AP designed by tailoring the IEEE Learning Object Metadata (LOM) to the satisfy the requirements of describing healthcare educational objects. The metadata schema is illustrated in Figure 2.3 and the new elements introduced in the scheme are highlighted in bold. A new category named *HealthCare Metadata* was added to the original nine categories composing the LOM standard. It provides information about the educational aspect of LOs in the field of healthcare education such as continuing education credits, patient and professional resources, and others [Smothers, 2004]. In addition to the elements provided in the educational cateogry of LOM such as resource type, interactivity level, its difficulty, the context where this resource can be used, and the typical age range, new elements in the *healthcare metadata* category provide further information related to the healthcare discipline such as *target audience*. Unique characteristics related to learning objects

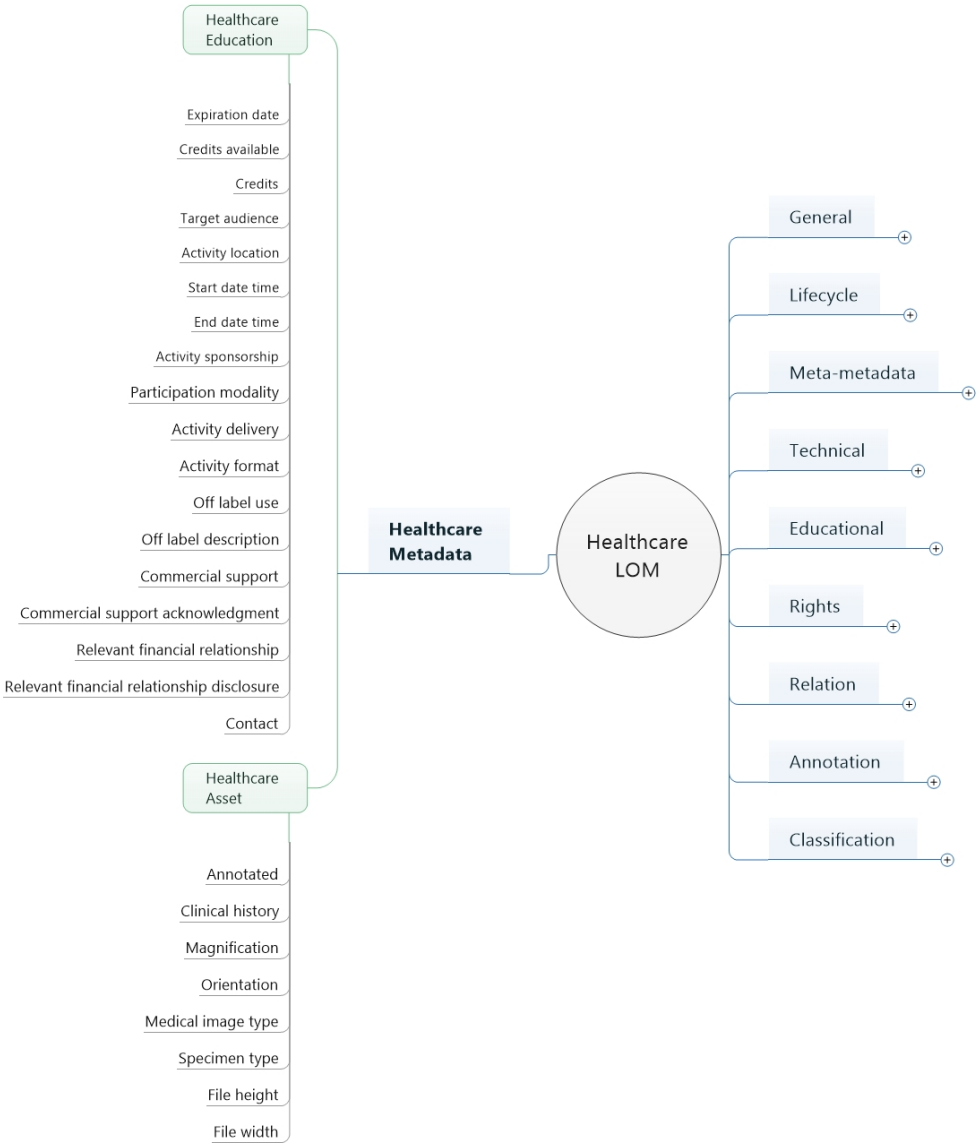


Figure 2.3: The HealthCare LOM metadata schema

in the field of healthcare is concerned with the cost of that LO; thus Healthcare LOM standard provides several data elements concerned with commercial and financial support. Furthermore, the healthcare learning objects are required to have copyright information. This information is stored in the *rights* category describes the licensing of the resource but does not state the privacy and confidentiality of the learner

information. Another important characteristic for any healthcare learning object is the trustworthiness of the source providing that LO. Hence, the source of the learning object can be stated using the *contact* element in the *healthcare education* category to allow providing more details about the provider of that resource.

2.3.2 mEducator

The mEducatror metadata schema was developed as part of a European funded project for developing best practice network launched to host medical educational materials from European higher academic institutions. The mEducator metadata schema was proposed in 2009, and the purpose of it was to enable ease of sharing, discovery, and reuse of medical educational content across EU higher academic institutions [Bamidis et al., 2009]. The mEducator metadata schema is considered an application profile that adopts the Dublin Core Element Set [Kunze and Baker, 2007]. New elements were introduced to the mEducator metadata schema to meet mEducator project's requirements, but not enough documentation is provided about how the AP was developed and which elements were extended or removed. The mEducator metadata schema is illustrated in Figure 2.4.

A resource can be described using mEducator metadata schema that consists of basic elements such as the ones provided in the *description* category. These elements exist for providing information about content description, technical description, and creation date in addition to other fields such as keywords, citation, metadata creation date and others. It also consists of elements that provide information related to the educational use of the resource it describes. For example, elements such as *educational objectives*, *educational outcomes*, and the *educational level* of a resource are all detailed in the *educational* category along with other elements such as, *learning instructions*, *assessment methods*, and *educational prerequisites*. An important part of the mEducator metadata schema that is related to the aim of proposing this schema is the *re-purposing* category. Since the schema was

developed with the aim of enabling sharing and reuse of resources between academic institutes, metadata elements that capture the re-purposing history of the resource are important. This category provides elements for storing information about the parent resource from which the current resource has been created if it is a reused resource. Also, mEducator provides the possibility of providing further information that certifies the content quality and appropriate classification of the resource.

The schema was implemented in both XML and RDF to ensure that the metadata was compliant with the principles of Linked Data [Mitsopoulou et al., 2011]. The values of the schema elements were restricted to controlled vocabulary sets provided in RDF too. For example, FOAF vocabulary [Brickley and Miller, 2012] is used to store information about the people involved in developing that resource. Also, SKOS [Miles et al., 2005] controlled vocabulary is used to restrict the entries to the *resource type* element. Furthermore, the *IPR* element used for describ-

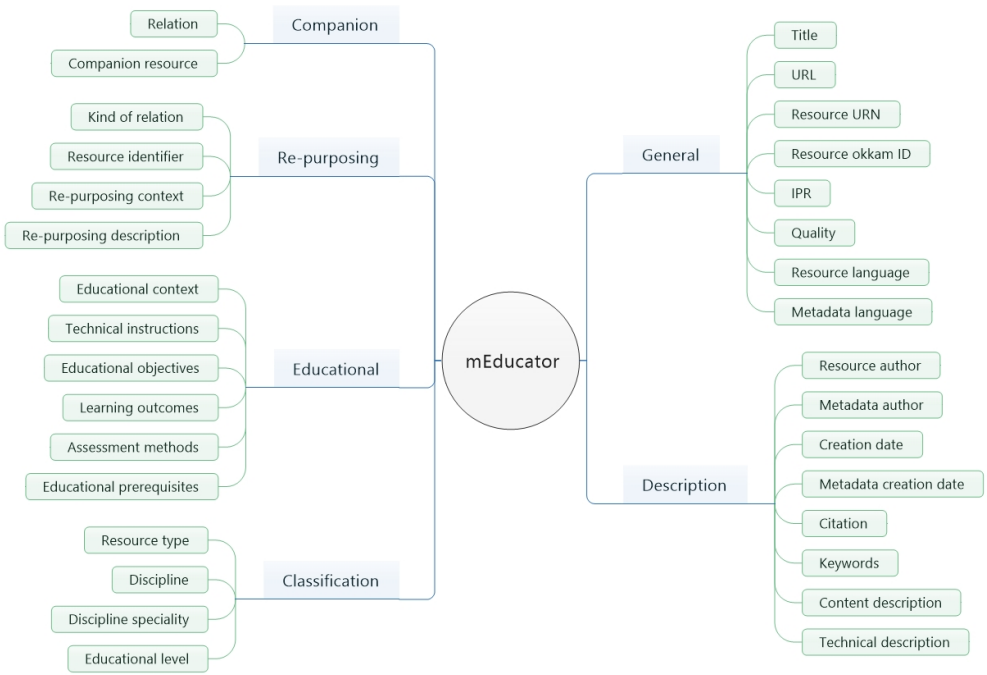


Figure 2.4: The mEducator metadata schema

ing the usage rights and licensing of a resource is represented by RDF vocabulary presented in the Common Creative licensing schema [Abelson et al., 2008].

2.3.3 HEAL

The Health Education Assets Library (HEAL) is designed to provide high-quality multimedia materials that are freely available to its users. HEAL project is a result of collaboration between numerous faculties, medical schools, the Association of American Medical Colleges (AAMC), and the National Library of Medicine (NLM). The HEAL application profile extends the Educause IMS metadata schema that was developed in collaboration with IEEE LTSC in 1998 [IMS Global Learning Consortium, 2006], and the HEAL application was made available beginning of June 2003 [Dennis et al., 2004]. The HEAL metadata schema is illustrated in Figure 2.5 and new elements were introduced that are highlighted in bold [Dennis et al., 2004].

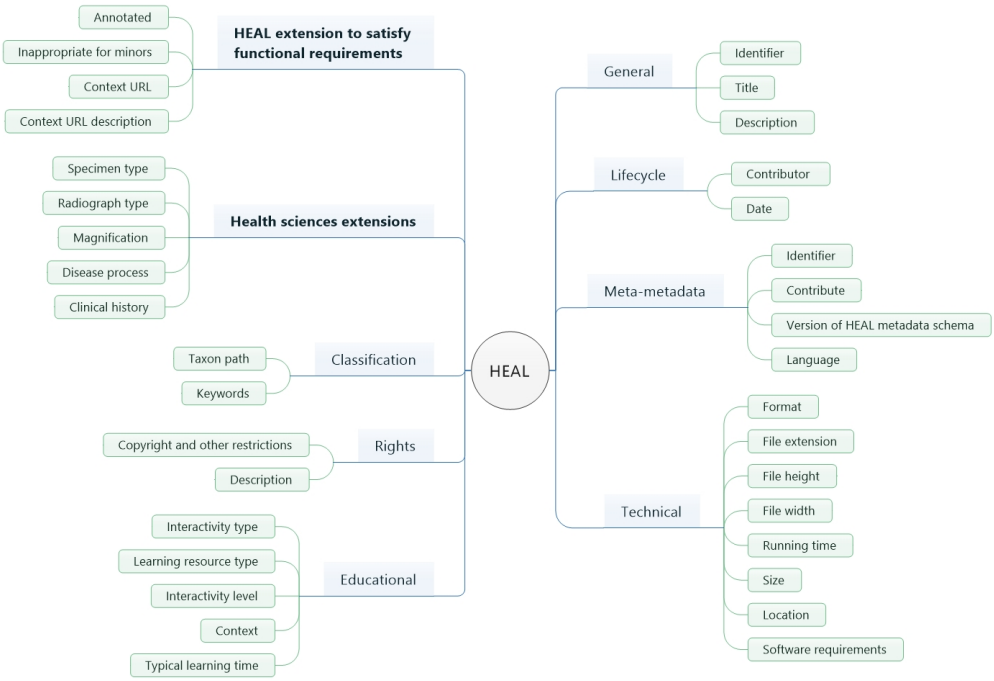


Figure 2.5: The HEAL metadata schema

The HEAL metadata schema provides a summary of about the content of the resource described in the *general* category of the schema. It describes some features about the educational use of the resource in the *educational* category such as the *typical learning time*. Further information can be provided in the *technical* category about the *software requirements* for using the resource described. Same as IEEE LOM, the *lifecycle* category captures the dates of creating the resource or modifying and the name of the contributor to this resource. In the *rights* category, HEAL schema provides information about assuring patient confidential content and stating clear intellectual property rights [Candler et al., 2003]. The contributor to HEAL collection of resources is responsible for providing details about the organisations or individuals involved in preparing the resources uploaded to it. One of the new elements added to the HEAL extensions is the *context URL* that provides a URL pointing to educational context in which the resource can be used. The context might be a course or a case that provide links to content with more relevant information that might benefit the learner.

The HEAL is implemented in XML format to ensure its interoperability and enable the use of controlled vocabulary to restrict the entries of some elements. Examples of such controlled vocabulary used in HEAL are MeSH vocabulary [Lipscomb, 2000], SNOMED CT [Stearns et al., 2001] and UMLS [Bodenreider, 2004].

2.3.4 NLM

The National Library of Medicine (NLM) exists in the campus of the National Institute of Health in Bethesda, Maryland, USA. It is the world's largest biomedical library that produces and maintains a large volume of printed collections and datasets in a broad range of topics to be searched by millions around the globe. The library has developed a metadata schema named NLM schema for describing its resources. The schema is an application profile that is based on DCMI schema and incorporates additional elements identified as requirements by NLM for publish-

ing its content. The NLM metadata schema is illustrated in Figure 2.6 and newly introduced elements are highlighted in bold.

The NLM schema is not developed for an educational purpose. It is used to index the resources in the library. As a result, the NLM metadata schema extends the simple DCMI metadata schema and adds few elements to enhance the organisation of the resources in the library [NLM, 2004]. It is not a complicated metadata schema and does not provide elements that is relevant to the healthcare field only in the *subject* category that is extended to include elements for describing the subject of

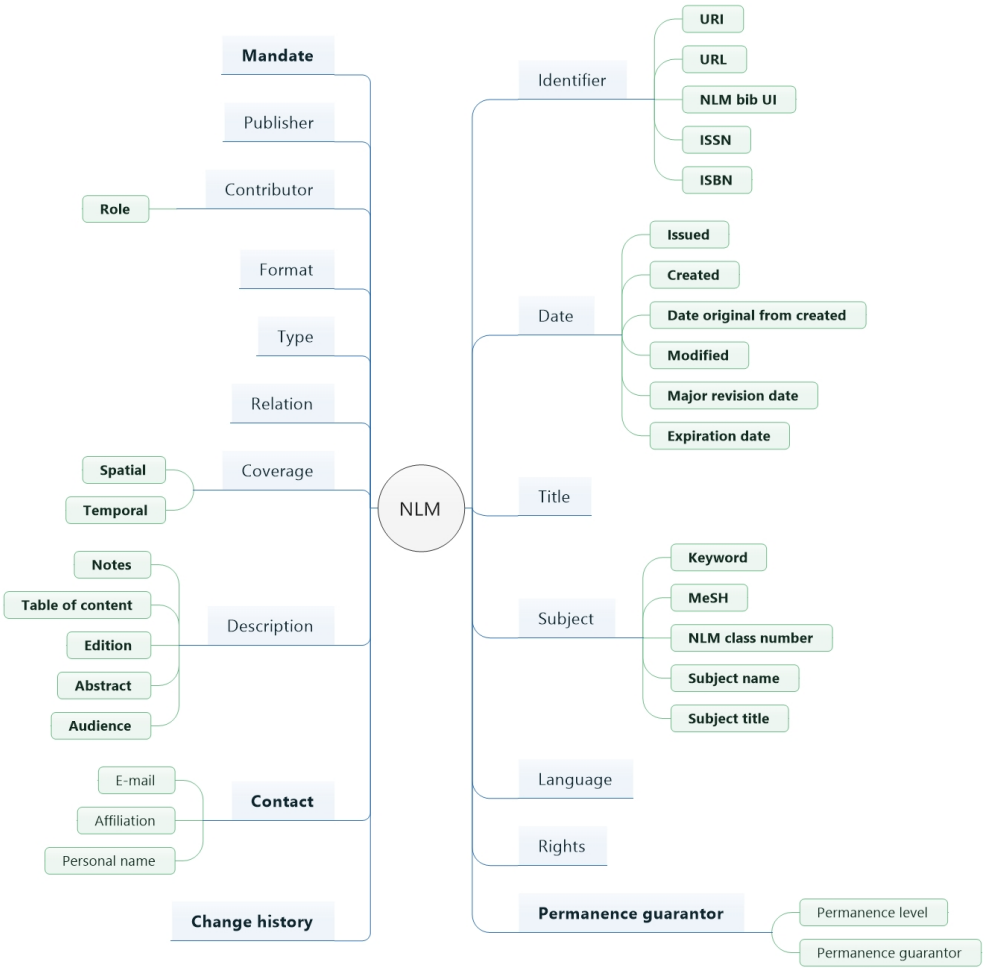


Figure 2.6: The NLM metadata schema

the resource based on the MeSH biomedical vocabulary. Furthermore, the *description* category is extended to include elements such as *notes* element that is used to provide free text space where any additional information about the resource might be added. This schema is extended from the DCMI schema (Figure 2.2) and uses similar elements to describe the general properties of the resource described. The NLM metadata schema focused on detailing information about the dates of revision and creation of the resources as noticed in the *date* category and the *change history* that differs from the DCMI schema. Another added element to the NLM schema is the *mandate* element that is responsible for supplying the name of a specific act or regulation if the resource requires legal instruments. The schema enriched the description of the *contributor* element with a sub-element to describe its *role*. Since the NLM schema is designed according to the specific requirements of a library, new elements were added to provide information about the guarantor of the resource described using the new category *permanence guarantor*. The information provided in this category state information that will assure the users of the library that the resource described will remain stable and available.

This NLM metadata schema is implemented in both XML and RDF/XML as the DCMI schema. Also, controlled vocabulary sets were used in this schema for restricting the entries of some elements. For example, the MeSH vocabulary is used for describing the *subject* element of the NLM schema and that helps in indexing the NLM resources.

All of the four metadata schemas (HealthCare LOM, mEducator, HEAL, and NLM) were applied in educational libraries in the medical education field. These schemas are developed as application profiles of existing well-established metadata standards such as IEEE LOM and DCMI. Table 2.1 summarises the main aims of developing these APs and the metadata standard they are extending. In addition to the main advantages and limitations of each AP.

Table 2.1: Summary of the four APs

AP	**Aims**	**Original metadata standard**	**Advantages**	**Limitations**
HealthCare LOM	Educational	IEEE LOM	Provide detailed information about healthcare specified educational properties	Complicated schema and requires trained librarians to complete
mEducator	Discovery and reuse	DCMI	Includes elements that detail the reuse of a resource, and support the use of different ontologies for filling its elements	Limited for use with educational materials published by European higher educational institutions
HEAL	Index and cataloguing	IMS	Support the use of several ontologies such as MeSH, SNOMED CT, and UMLS to fill its elements	Complicated and require trained librarians to complete
NLM	Index and cataloguing	DCMI	Simple schema that support using MeSH concepts for categorising the resources it describes	Designed to index the content of the National Library of Medicine only

2.4 Publishing Data on the Web

In the 1980s, the internet started to emerge after the work of Tim Berners-Lee on the World Wide Web. He theorised that the hypertext documents can be linked

via protocols making one working system [Couldry, 2012]. Since the mid-1990s, the Internet has had a revolutionary impact on the world. It caused a change on the communication sector with the emergence of electronic mails, instant messaging, and Voice over Internet Protocol (VoIP). Not only in communication, the Internet had become the preferred platform for data publishing in all the fields such as education, support, finance and media. Furthermore, it advanced to be the platform for data sharing and exchange after the spread of the internet over the world. From the beginning of this revolution, several methods and techniques have been developed for publishing data on the web. This section presents a brief history about such methods with the focus on describing the Linked Data practices that are recently

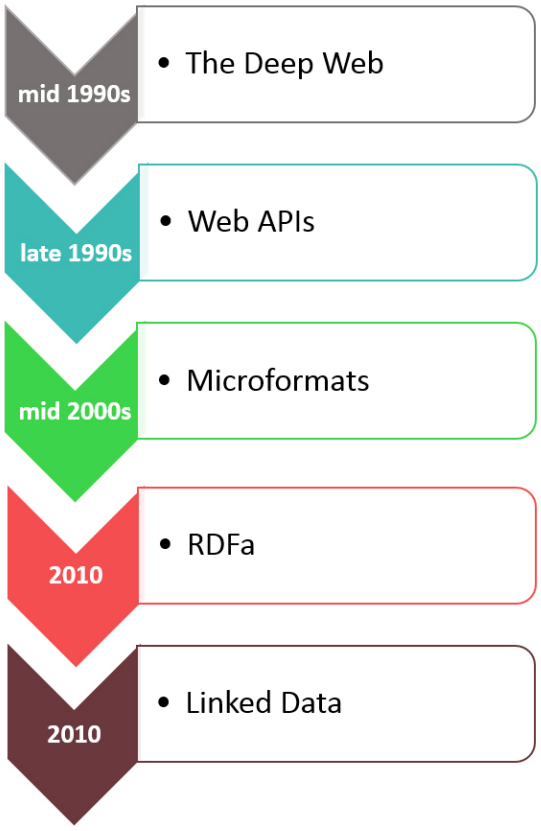


Figure 2.7: The history of web data publishing

becoming very popular in web data publishing. The advances in the field of web publishing is shifting how the web is formed. This thesis is concerned with collecting data from the open web and publishing it as Linked Data which is considered the latest technique for publishing data on the web.

2.4.1 The Deep Web

This term started to appear in the early days of the web in the mid-1990s. It is used to identify data that are hidden from search engines and can not easily be accessed [Chang and Cho, 2006]. The deep web provides access to a database using a query form designed for searching the content of that database. This option for publishing data is still common these days because it is not costly to implement for the data owners, and it provides value to the final users. Standard search engines create their indices with the help of crawlers and spiders that navigate the surface of the web. Some data sources publish their content on the deep web that is not accessible by spidering or crawling, thus such content is only searchable using a query form that produces results dynamically for each query initiated. The size of the deep web is massive and it is estimated to be 500 times more than the searchable content on the web [Bergman, 2001]. Technologies such as OpenSearch or the BrightPlanet were developed to overcome the issue of searching the deep web [Ghanem and Aref, 2004]. The work presented in this thesis is focused on organising the open data found on the web and is not concerned with the data published in the deep web. The following sections summarise newer methods that ease the retrieval and publishing data on the web.

2.4.2 Web APIs

With the widespread use of the internet, much efforts were put into standardising the methods of data exchange over the web. The web Application Programming Interfaces (APIs) emerged in the late 1990s to improve the data exchange on the

Web [Ceri et al., 2013*b*]. In the deep web, the query forms result in data published as HTML documents shared with the users. After the proposal of the XML standard for web data encoding, the idea of web services has emerged and was mainly centred on using XML for data exchange. Sometimes Web APIs is used as a synonym for web services. A web API is a programmatic interface which provides endpoints that define request-response messaging service for a web application via the common means of the web that is HTTP [Yee, 2008]. The request and response messaging are usually represented in XML until the emergence of the JSON format that is widely adopted on the web. Also, new types of web services emerged in the late 2000s based on the representational state transfer (REST) paradigm known as RESTfull services [Richardson and Ruby, 2008]. Over the years, the preferred technical solution has changed, but the number of Web APIs available have steadily increased. ProgrammableWeb[1] is the largest Web APIs directory, and it reports on having more than 14,000 web APIs in February 2016, and the numbers are growing exponentially. Even large social networks such as Facebook and Twitter published Web APIs to access their content. The data behind the Web APIs are inaccessible to search engines. It results in a new type of Deep Web because data publishers prefer to have control over their data and provide control over how it can be used. Hence, *mashups* emerged as a practice for combing data from multiple APIs thus introducing a new type of web services [Benslimane et al., 2008].

2.4.3 Microformats

Starting from the mid 2000s, movements for publishing data openly on the web have emerged. The focus was to provide meaningful data on the web for users and application to consume. Hence, the microformats approach of publishing HTML content that is marked up with metadata to convey its meaning has emerged [Ceri et al., 2013*b*]. Using HTML and XHTML, web pages were annotated with machine-

[1] http://www.programmableweb.com/

processable tags to provide information about the web pages content [Khare, 2006]. For example, marking a web page about an event with *hCalender* microfromat allows search engines such as Google to read the date, time, and location of that event, which is linked to GoogleMaps. That process provides a rich search result for the users using the web. Several microformats have been deployed for annotating the web content. For example, *hCard* is used to annotate data about people and organisations, *hReview* is used to mark up product reviews, in addition to more microformats developed and used on the web. This method provides data that is machine-processable, and that helps in providing richer search results and better representation of the data published on the web [Mika, 2008]. The method had a positive effect on the way users navigate the web. It helps in connecting related information on the go using one popular microformat named *RelTag*. It reduces search efforts needed for the users to find the relevant information they want.

This method is considered a major turning point for web data publishing and it has paved the road for further methods proposed for organising the web. The microformats had some limitations that had to be overcome. For example, for each microformat deployed on a web page, it needs a specific parser to read and process it. Moreover, the microformats can not be generalized. They are proposed for specific usage and with specific attributes that might be used in other microformats causing a collision between the attributes tagging a web page [Ceri et al., 2013*b*]. Later methods proposed for enhancing web data considered annotating the web but with techniques that takes into consideration solving the limitations of the microformats.

2.4.4 RDFa

The RDFa method emerged after the vast adoption for the microformat method for annotating the web. It was proposed to address the limitations of the microformats. The RDFa provided a specification for the attributes annotating the web pages with marked up tags making it machine processable [Adida, 2008]. RDFa pro-

vides the syntax for the tags used to mark up HTML/XHTML web pages [Adida et al., 2008]. Having such standardisation for describing the web pages, independent initiatives started to standardise the terms used to describe these attributes such as the *Schema.org*. Using a unified vocabulary can be beneficial for parsing and intermixing web pages that uses the same vocabulary. The *Schema.org* is a joined initiative between Google, Yahoo!, and Bing that aimed at specifying a broad vocabulary that is parsed by these search engines when used to publish web pages [Barker and Campbell, 2014].

The RDFa has been widely adopted in the last years. In 2010, Facebook introduced OpenGraph[2] that is a protocol that allows web pages to be integrated into Facebook. It proposed attributes and vocabulary set that is used with the RDFa to enable the publisher to describe their web pages making them accessible across search engines, social networks, blogs, and so on. The vast adoption of such data publishing techniques changed the idea about web data publishing. Nowadays, the web can be considered as pieces of data related to each other instead of collection of documents with typed links connecting them.

2.4.5 Linked Data

The previous sections have explained traditional techniques for accessing data published on the web [Ceri et al., 2013*a*], such as the deep web (section 2.4.1) and the web APIs (section 2.4.2), and modern techniques for enriching the content of web pages such as microformats (section 2.4.3) and RDFa (section 2.4.4). Linked Data practices emerged in an intention to bridge the gap between these two types of techniques. It provides the principles for publishing data on the web that is rich with semantics and interlinked with external dataset [Bizer, Heath and Berners-Lee, 2009]. The idea of Linked Data is aligned with the idea of the semantic web. It has changed the web structure from a space of linked documents to a space of linked

[2]https://developers.facebook.com/docs/sharing/opengraph

data that is named the "Web of Data" [Bizer, Heath and Berners-Lee, 2009]. The concept of Linked Data is to interlink data on the web and publish it in a machine-readable format using standard web technologies such as URIs and RDF [Heath and Bizer, 2011].

In 2006, the method of publishing Linked Data one the web was provided by following these guideline [Berners-Lee, 2006]:

- Use Uniform Resource Identifiers (URIs) to name things
- Use Hypertext Transfer Protocol (HTTP) URIs to look up things
- Use standards such as RDF and SPARQL to provide useful information in response to URI lookup
- Include links to other URIs to allow more things to be discovered

Two fundamental technologies that Linked Data relies on are URIs [Berners-Lee et al., 2005] for identifying any resource on the web, and the HTTP [Fielding et al., 1999] protocol that is used to dereference these URIs. A full glossary of terms used in Linked Data practices and its associated vocabulary are maintained and explained [Hyland et al., 2014]. When Linked Data are published on the web, it is represented using RDF format [Cyganiak et al., 2014]. The RDF technology is critical to the Web of Data. It provides a graph-based data model to structure and link resources described on the web. Its job is similar to how HTML is used to structure and link documents on the web. The RDF model is a method for encoding data as triples in the form of subject-predicate-object. Thus allowing any two resources to be linked using a relation defined as a predicate. The subject part of the triple must always be a resource identified by a URI while the object may be another resource or a literal such as a string or a date. The predicate specifies how the subject and the object are connected and is identified by a URI. The RDF triple can be expressed as a graph shown in Figure 2.8.

Figure 2.8: RDF graph example

The data is represented in RDF, but other modelling languages are used to create vocabularies or often called ontologies [Horrocks, 2008] that are used to define the meaning of the data. Such modelling languages are RDFs [Brickley and Guha, 2014] and Web Ontology Language (OWL) [Van Harmelen and McGuinness, 2004]. They are used to define terminologies in specific domains that can be used to represent the relations between the resources. For example, the Linked Data sentences represented in Figure 2.8 can be rewritten using such vocabulary that will enhance the readability of the statements and can be represented in Figure 2.9.

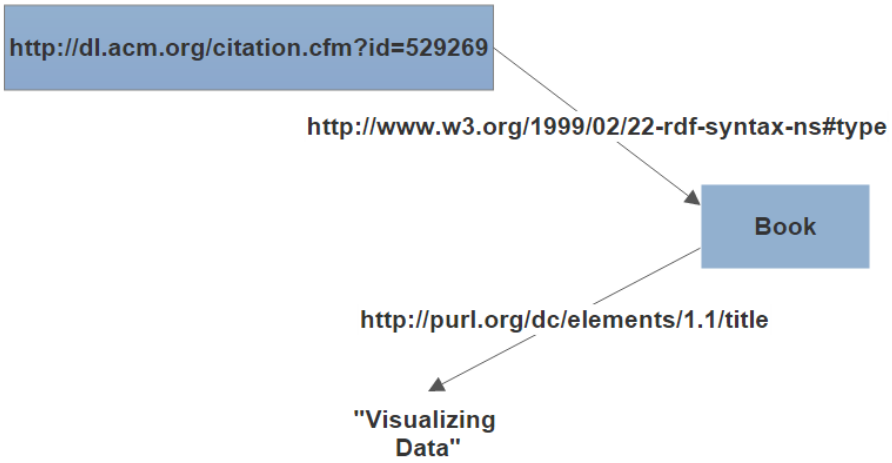


Figure 2.9: RDF graph example using controlled vocabulary

Although anyone can create vocabularies and publish them to be used with RDF triples [Bizer, Heath and Berners-Lee, 2009], it is preferable to reuse exist-

ing vocabularies if possible [Bizer, Heath and Berners-Lee, 2009; Heath and Bizer, 2011]. Using well-known vocabulary to describe data makes it possible for various applications to consume and understand. Once the data are described with RDF triples and the appropriate vocabularies have been used, it can be published using an RDF triple store [Heath and Bizer, 2011]. These RDF triple stores are accessed and queried via the SPARQL query language [Prudhommeaux et al., 2008].

Despite the fact that the Linked Data emerged few years ago, the broad adoption of these principles has been very successful. The Web of Data has been growing exponentially since then [Heath and Bizer, 2011]. The W3C Linking Open Data (LOD) project represents the datasets published using Linked Data in the LOD cloud [3] and statistics about these datasets are periodically reported to compare the growth of Linked Data on the web [Cyganiak and Jentzsch, 2011]. One of the most popular examples of Linked Dataset is the DBpedia project [Bizer, Lehmann, Kobilarov, Auer, Becker, Cyganiak and Hellmann, 2009] that is used to build a linked dataset of Wikipedia content automatically. Although the process of publishing Linked Data is not a trivial process, it provides a mechanism for integrating data over the web and facilitate its discovery. With the RDF model acting as a unifying data model that represents resources in URIs and using standard vocabularies that ease the integration of data from different datasets [Heath and Bizer, 2011], the future of the web looks promising. Hence, this research is concerned with exploiting the Linked Data practices for aggregating and integrating open educational content published on the web.

2.5 Linked Data in Education

Significant research efforts have focused on applying semantic web technologies in the learning domain that covered different research areas such as applying semantics to recommendation systems, intelligent tutoring systems, and enriching the meta-

[3] http://lod-cloud.net/

data of learning objects [Devedžic, 2006; Sampson et al., 2004; Naeve et al., 2006; Bertini et al., 2011; Carmichael and Jordan, 2012; Tiropanis et al., 2012]. Although applying semantic web technologies to support the discovery and delivery of educational content on the web have been effective [Carmichael and Jordan, 2012], such research efforts have resulted in another problem. Several datasets exist on the web that are enriched with semantics, but they are hard to integrate together due to the lack of defined terminology that annotates educational content of different domains [Tiropanis et al., 2012]. Moreover, the idea of producing datasets that are ready to reuse and share on the web was not popular, and few datasets were prepared to accommodate this need [Shadbolt et al., 2006; Hendler, 2008].

With the emergence of the Linked Data practices (section 2.4.5), large volumes of data were available and ready for sharing and reuse due to the unified standards suggested in the Linked Data principles. In the learning domain, the efforts of producing open data were significant resulting in having several number of datasets and applications for providing open educational content for the community. As a result, projects have been developed to maintain catalogues of such initiatives. For example, Linked Universities[4] lists some educational institutes that release their datasets as Linked Data. Furthermore, initiatives such as Linked Education[5] and Linked Education Cloud[6] catalogue educational datasets published in Linked Data. Also, competitions like the LinkedUp Challenge[7] encouraged the research towards applying Linked Data in education by looking for interesting applications that are concerned with analysing or integrating web data for educational resources.

The Linked Data are applied in different research areas in education and it appears to have some effect on applications that support learning and teaching. It is employed for supporting interoperability of learning objects, and applying novel ways for collaboration and personalisation of learning. Moreover, Linked Data can

[4]http://linkeduniversities.org/
[5]http://linkededucation.org/
[6]http://data.linkededucation.org/linkedup/catalog/
[7]http://linkedup-challenge.org/

be used for data integration and maintaining a well-formed metadata for learning objects [Tiropanis et al., 2012]. A full review of Linked Data proposals in the learning domain is presented in [Vega-Gorgojo et al., 2015] that analyses existing research work in the literature. Despite the fact that applying Linked Data in education can be challenging [Dietze et al., 2013], yet it has been reported about the opportunities it provides for open and distance learning [dÁquin, 2012; Vega-Gorgojo et al., 2015]. Existing research efforts focused mainly on the interoperability of learning resources as this is the principal goal of Linked Data. Furthermore, enrichment of educational content was the focus of application in the research areas of personalisation of learning and contextualised learning. Some research efforts had focused on applying Linked Data in game-based learning and mobile learning. In the rest of this section, a discussion of some of the existing research in these fields is presented to highlight the aim of applying Linked Data in educational applications.

One interesting research is an application that used datasets from Greek DBpedia to build a web game [Bratsas et al., 2012]. The game consists of several quizzes that are used for educational purposes with native speakers. The application has been developed with internationalisation in mind. The recent efforts of DBpedia projects enabled the internationalisation of its content starting with the Greek DBpedia [Kontokostas et al., 2012]. Thus, the application can be applied to different DBpedia language editions and the learning resources can be interoperable between different languages. Another research focused on transforming lecture notes into a document that can be interactively browsable by students [David et al., 2010]. A collection of lecture notes in maths were semantically enriched and published as Linked Data that is exposed and ready for integration with other external resources. Also, Linked Data have been used to publish open data about educational institutes such as the research effort conducted for publishing statistical data about Italian universities in [Pirrotta, 2010].

In the research area of personalising learning, Linked Data have shown

promising results when incorporated in applications such as adaptive learning systems. Linked Data principles can be used for modelling user profiles after mining data collected from social web applications [Abel et al., 2011]. Other research efforts investigated the automatic delivery of Learning Objects (LO). The adaptive delivery was implemented by enriching the LOs metadata with external web data in [Yoosooka and Wuwongse, 2012], and the users' profiles were enriched with the same web data to enable automatic retrieval of LOs that matches the needs and desires of the users. Besides personalisation of learning systems, Linked Data was applied in social learning applications that encourage social engagement with learning such as the work presented in [Jeremić et al., 2013]. It builds a personalised learning environment for students to help them learn by providing them with resources of their interest. Also, Linked Data has been applied to different platforms such as mobile application to enable browsing Open CourseWare (OCW) from various universities [Piedra et al., 2012]. Several OCW repositories were integrated into one dataset that can be accessed via a mobile application and the application personalised the browsing experience based on the user interactions and the data that can be collected using the mobile such as *location*. A lighter weight application is presented in [Robinson et al., 2012] that supports the idea of social learning using e-books. Linked Data is used to leverage the user-generated content for representing annotations of the e-books content and modify the e-book display according to the user interest as they state at the beginning of the learning process.

It is noticeable that the research covered in the literature on applying Linked Data in education had spanned over different educational sectors. Linked Data have been implemented to enrich the data resulted from using a platform for childhood education and care. Tools have been designed to enhance educational resources recommendation, nutritional monitoring, and health monitoring services based on Linked Data [Alonso-Roris et al., 2012]. External datasets were used to supplement and extend the information resulted form this educational platform to enhance its

services. Other projects targeted the management of learning materials available on the web. SemUnit project initiated by French higher education institutions exploited Linked Data to integrate French repositories that contain learning materials of high-quality for different domains [Isaac et al., 2012]. Ontologies have been incorporated for enriching the metadata of learning materials in these repositories such as FOAF for describing persons and organisations and SKOS for describing controlled vocabularies in metadata elements. The metadata of the learning materials in these repositories are described in SupLOMFR schema[8] that is a LOM application profile dedicated to the French higher education institute. Similar research efforts have been applied in the field of biomedical education for integrating learning resources provided by two open educational repositories: the PubMed[9] and the OpenLearn[10] [Dietze et al., 2012]. The metadata of these learning resources is described using mEducator schema [Mitsopoulou et al., 2011] explained in the previous section (section 2.3.2). The project provides an infrastructure for educational data and services integration. As part of the mEducator project an application has been developed named *MetaMorphosis+* that provides an environment for sharing linked educational resources [Hendrix et al., 2012]. In another application domain, [Sicilia et al., 2011] discussed the advantages of exposing the content of the Organic.Edunet portal that is a federation of learning repositories in the domain of organic agriculture. It is apparent from the literature that enriching educational content with external data is one of the important usages of Linked Data. Such enrichment can be applied to improve the classification of learning objects in one repository such as the use of DBpedia categories for improving the automatic classification of Learning Objects in a university library [Lama et al., 2012].

One of the advantages of applying Linked Data to education is its ability to provide a well-formed metadata for learning resources on the web. Some re-

[8]http://www.sup.lomfr.fr/index.php/Accueil
[9]http://www.ncbi.nlm.nih.gov/pubmed
[10]http://www.open.edu/openlearn/

search efforts focused on enriching the metadata description of particular types of resources such as videos. The poor metadata structure of such resources reduces its discoverability. Therefore, the work presented in [Fernandez et al., 2011] focused on extracting, structuring, and interlinking video lectures published by different educational institutions. This resulted in one repository of integrated videos extracted from websites and *YouTube* channels. Another research proposed a solution for enhancing the discoverability of videos that lack sufficient metadata and proposed a semantic video search engine named "yovisto" [Waitelonis et al., 2010]. It presented an exploratory search engine that expands the search query with terms extracted from DBpedia. For an improved discoverability, both the metadata and the query can be extended and enriched with Linked Data as applied in this research [Yu et al., 2012], and that builds a platform for browsing video annotation. This technique was applied on a repository of videos provided for the history course at the Open University. The video resources are annotated and semantically enriched with external web data to enable its browsing.

Another usage for the Linked Data in education that has been investigated in the literature is the possibility of authoring educational content from existing open educational resources published in Linked Data. Text and video information contained in blog articles are enriched in [Ruiz-Rube et al., 2011] in order to provide a set of resources that are semantically annotated. This research proposed a tool that automatically selects related resources from the blog entries to be delivered to students running a learning activity. Using Linked Data published on the web provides the opportunity to overcome the costs of building and creating content on the web. For example, the work presented in [Ruiz-Calleja et al., 2012] proposed an approach based on Linked Data that allows the integration of educational ICT tools on the web. One successful experience of publishing open data is provided by the UK Open University, and it is considered a blue print for other organisations to open up their data and enables its sharing and reuse [Zablith et al., 2015].

2.6 Gaps in the Literature

One of the main aims for developing and using metadata schemas for describing resources on the web is to enhance the search and retrieval of resources on the web when searching. New advancements in the field of e-learning standardisation have emerged to facilitate better description of educational resources in online libraries. Formats such as XML and RDF have been used for implementing metadata schemas making it machine-readable. Furthermore, such formats are capable of utilising open linked datasets and ontologies to enhance the description of the metadata making it machine-understandable. Despite all these advancements, there are some limitations in the existing metadata standards resulting new metadata schemas and APs to be proposed. This research is not intended to solve this issue and propose an ideal standard that fits all the resources, but it addresses the following gaps noticed in e-learning metadata standards that affects the discoverability of resources on the web.

- Metadata are designed to describe a specific type of e-learning resources. Usually, metadata are designed to organise the content of a library or a repository. Therefore, the types of the learning resources are limited to one or two types when using metadata schemas such as mEducator, HEAL, NLM.

- The values of metadata elements can be restricted to one controlled vocabulary sets by assigning one vocabulary set for one data element, such as the use of MeSH vocabulary in NLM. It is noticed that the use of controlled vocabulary with metadata schema elements is focused on indexing or categorising the resources in online libraries, thus providing a rigid metadata that is restricted to external data.

- Some metadata schemas are complex and require qualified staff and librarians to fill and manage. The mEducator metadata schema is one of the schemas

developed for describing medical learning objects that is built on top of DC metadata and used to describe educational content from two online libraries. The metadata schema is not simple and it focuses on describing the resources for re-purposing in education. This metadata schema is developed as part of mEducator European project. The second phase of the project proposed a social network that is built on top of the mEducator schema named metamorphosis+[11]. The content of this network at the time of writing this thesis is not of a high quality and the resources uploaded by the users of this network are not complying to the metadata schema when describing the resources they share on this network.

This research intends to address the gaps of using metadata standards in a specific domain that is the medical education domain. A metadata schema is developed as an application profile of the DC metadata schema to describe existing educational objects published on the web. The main aim of this metadata schema is to enhance the discoverability of educational objects and add semantics to its metadata that enable its linkages. Hence, an exploratory study with people involved in the domain of medical education will clarify the requirements and the needs of web users when it comes to searching for educational content on the web. The big picture for the proposed metadata schema is summarised in the following aims.

- It accommodates different types of educational objects published on the web. The proposed metadata schema should bridge the gap between traditional educational objects such as articles and Web 2.0 content that is utilised by web users for learning such as *YouTube* videos and blogs.

- It can be enriched with external ontologies and external datasets without restricting the value of its elements with a controlled vocabulary set. This is possible with the use of Linked Data for implementing the metadata schema

[11] http://metamorphosis.med.duth.gr/

and the methods of enriching its content. Any ontology can describe the metadata elements in the proposed schema by linking it to the ontology concepts using specific elements proposed in the metadata to enable such enrichment.

- It should be simple and interoperable to accommodate different metadata schemas extracted from the web. The metadata schema should be easily mapped to existing metadata schemas used to describe learning resources on the web in order to be filled automatically by a system and not the users. Automatic harvesting and extraction of data from the web can be developed to increase the quality of the metadata records. Furthermore, enriching the metadata with external data sources can be automated to enhance the metadata records and link them to external data that adds semantics to its content. Thus, enhance the discoverability of the resources.

The proposed metadata schema is designed based on the recommendations deduced from the exploratory study conducted with users from the medical education domain. The full potential of the proposed metadata schema is reached if incorporated with a system that will enable the use of the features introduced in this metadata schema. The rest of this research explains the process of designing, developing, and testing the metadata schema proposed and the system designed to enable its usage.

The previous section discusses different projects applied in education for exploiting Linked Data to improve discoverability, delivery, integration of educational content. Some research focused on personalising education, others were concerned with building intelligent learning systems, or recommender systems. This research is concerned with applying Linked Data for integrating educational objects from distributed sources. Several research projects covered in the previous section were concerned with integration of educational content. The area of application differs from one project to another. One research projects was applied for integrating two

repositories that provides educational resources for medicine and healthcare that is the mEducator [Mitsopoulou et al., 2011] project. It is proposed to provide a platform for sharing and exchanging of medical educational content. Other research focused on integrating repositories managed by one party such as SemUnit [Isaac et al., 2012] project initiated by the French higher education or the organic.edunet project [Sicilia et al., 2011] for integrating repositories in the field of organic culture. Also, research efforts have been applied for integrating videos published on *YouTube* channels [Fernandez et al., 2011].

All the research efforts applied for integrating resources were limited to specific datasets or specific repositories hosting similar types of educational content. There is no effort, to the best of our knowledge, that aims at integrating diverse types of web data into one repository. The main aim of this research is to bridge the gap between diverse types of content that is published on the web and can be used for education. The goal is to build a dataset of articles, blogs, and videos, harvested from online libraries and Web 2.o platforms.

2.7 Summary

This chapter presented the research background and related work in metadata standardisation and Linked Data applications in education. It has described two of the well-established metadata standards applied in e-learning. Furthermore, it has examined in detail the metadata schemas resulted from the research efforts related to managing healthcare educational resources. On the other hand, this chapter has presented an overview of the advancement of the web data publishing techniques. Followed by a detailed explanation of the Linked Data, what is it, and how it can be adopted for publishing data on the web. As this thesis is concerned with applying Linked Data for integrating educational medical content on the web, this chapter has examined research efforts in the literature that is concerned with applying Linked

Data in education.

In conclusion, a part of this chapter that is concerned with metadata standards has addressed the research objective **O1**: "Identify existing metadata schemas that are already used for describing EMOs in specialized medical educational libraries". By addressing this research objective, this chapter partially answers the research question **R1**: "What are the current metadata schemas used when publishing medical educational content on the web and what are the essential elements from the user perspective when searching for such content?". The gaps identified in the literature and the outcomes of examining the related work have provided an initial guidelines for developing the solution for the problem investigated in this research. Further exploratory research is conducted in the following chapter focusing on discovering the challenges and preferences of the medical educational community.

Chapter 3

The Exploratory Study

3.1 Introduction

The problem identified in this thesis is concerned with aggregating and integrating educational medical objects from distributed web data sources. It also involves the issue of accessing and retrieving objects from such databases as stated in the problem definition (section 1.1). Such problems are affecting the users of the web who search for educational content. Engaging the user in designing a solution to solve the problem being investigated improves the quality of the final results. Hence, an exploratory study that helps to elicit the frustrations of web users when searching for educational content has been conducted. Such study can produce the functional requirements needed to develop the solution. The preferences and attitudes of the community can be deduced, and that affects the process of designing and developing the solution proposed.

The exploratory study, reported in this thesis, is concerned with learners in the field of medical education since the problem identified was restricted to the medical education domain as a proof of concept in this research. Although the community interested in learning about medicine and healthcare is wide, the targeted domain of this exploratory study was narrowed down to students and educators

involved in medical education. Understanding the current practice of these learners and identifying the difficulties they might encounter, reveal valuable information to support the design of the solution. As a representative sample, this exploratory study was conducted with the cooperation of Warwick Medical School (WMS) students and educators. In this chapter, the exploratory study aims to examine the Knowledge, Attitudes, and Practices (KAP) for a sample population of learners in the field of medical education regarding using the web for education. The KAP survey conveys the purpose of this study and serves as a diagnosis tool for the domain studied [Kaliyaperumal, 2004]. The results of studying a sample group can be generalized to represent the opinions and attitudes of learners in the domain of medical education [Launiala, 2009]. The KAP survey acts as a research tool that applies a quantitative method of collecting data via questionnaires. The information collected from this survey reveals the opinions of its participants and it is based on declarative statements provided by the targeted population in the surveys [Médicins du Monde, 2011].

The findings discovered in the KAP study suggest a set of recommendations which support the design and development of the proposed solution. The outcomes of this chapter represents the preferences and attitudes of the targeted domain towards proposed techniques incorporated in the solution. Additionally, the background research revealed gaps in existing metadata schemas designed in the medical education domain. Hence, the KAP survey focus was to view the opinions of a sample from the medical education community on particular issues recognized in the background research. The gaps listed in the background research and the recommendations resulting from the KAP survey contribute to the functional requirements elicited for developing the techniques and methods detailed in Chapter 4 and Chapter 5.

3.1.1 Chapter Objectives

This chapter aims to realize the research objective **O2**: "Identify the search practices and challenges faced by students and educators in the medical education field when searching for educational content on the web including Web 2.0 sites and online academic libraries". The process of addressing this research objective helps in suggesting recommendations that support the development of the solution for the problem identified in this research.

3.1.2 Chapter Outline

The rest of this chapter is organized as follows. The objectives of the exploratory study are elicited to clarify the main aims of the KAP survey. Followed by a detailed explanation of the study methodology applied to conduct the KAP survey. The survey results are summarized next and followed by a detailed discussion of the significant results. Finally, the list of recommendations suggested after analysing the answers of the population studied.

3.2 The Study Objectives

The KAP survey aims to report on the knowledge, attitudes, and practices of the WMS students and educators when searching for educational content on the web. The objective of this exploratory study is to analyse the answers of the respondents and suggest recommendations that support addressing the gaps identified in the background research. The KAP survey was conducted with both students and educators of the WMS. Including two different sections of respondents in this study enables the study to reach a more representative sample of the medical education domain studied. The main objectives of the KAP study are summarized as follows.

1. To identify the challenges encountered by the respondents when searching web data sources for educational content.

2. To identify the current practices of students and educators when searching the web for educational content.

3. To identify the attitudes of the respondents toward adopting some proposed techniques that might enhance the search process.

4. To rank the importance of the findings discovered from the respondents' answers concerning their attitudes and practices on existing or suggested search practices.

The KAP survey distributed when conducting this exploratory study is straightforward and gathers information about the targeted population studied. It focuses on gaining knowledge about how acquainted the respondents are with the concept of Open Educational Resources (OER) since this research is directed towards open data on the web. The rest of the KAP survey focuses on collecting information about the respondents' frustrations, current practices, and preferences for improving their experience while searching for educational content. More comprehensive understanding of the medical education community helps in forming the functional requirements that guides the process of designing and developing a solution to the problem identified.

3.3 The Study Methodology

For the sake of reaching a large number of respondents from both the students and educators in the WMS, the KAP survey was conducting online instead of in-person. The staff directory of WMS[1] provides details about the staff including their names, emails, and job titles. Emails were sent to the academics and professors of the WMS provided in this directory. As for the students, a link to the online survey was shared via the official social networks channels shared on the WMS website[2].

[1] http://www2.warwick.ac.uk/fac/med/staff/
[2] http://www2.warwick.ac.uk/fac/med/

Although the targeted population was large, the participation in the survey was voluntary. Hence, the response rate was relatively low, and the final number of respondents was less than expected. However, the information collected from the KAP survey was sufficient to identify the knowledge, attitudes, and practices of learners when searching for educational materials in the medical education domain.

Targeting two different section of respondents at the WMS was beneficial. Both the students and educators in the WMS school have had previous expertise of studying or teaching in other institutions that span over the UK or the world. It helped to gain a better perspective on the different practices and attitudes of the two sections. The KAP survey distributed to both sections were the same except few questions (3 questions) used to report on the characteristics of the respondents participating in the study. For example, students were asked about the course they were studying, while the educators were asked about what courses they taught. The rest of the survey (7 questions) focused on gathering quantifiable data that can be reflected on the targeted domain. The questions are summarized in the following categories.

- Knowledge: Are participants aware of the OER initiative?
- Attitude: What are their frustrations when searching the web? If introduced with new criteria for filtering the search results, does the respondents think it will be beneficial? Moreover, what new criteria do they think will enhance their searches?
- Practice: How often do they search the web for educational content? Where do they search? Also, do they use any filtering criteria for editing the search results?

More specific questions for the students' section were included in the survey aiming to capture their tendency to share knowledge with their peers. Nowadays, the revolution of the social networks has affected the practices of students learning and

such information will be beneficial for future plans of this research. WMS presented their interest in this research and the necessary approval was sought for conducting the study with its students and educators. Full ethics approval was awarded by the University of Warwick Biomedical & Scientific Research Ethics Committee (BSREC) reference (REGO-2013-060). The surveys were hosted online and links were emailed to the educators and shared with students via the social networks of WMS. The questionnaire rubrics stress the confidentiality and anonymity of the responses and explain the use of the data collected in the research. Participants proceed and conduct the survey if they agree on the information provided in the email which act as an informed consent of participating in this research. The results of the KAP survey are detailed in the following sections, and detailed discussion of these results produced the recommendations from the targeted domain.

3.4 The Study Results

The two sections involved in this study are students and educators in Warwick Medical School. The surveys were shared with large number of students and educators to have a reasonable number of respondents considering the poor replies of online surveys in general. The final number was 136 respondents: 62 students and 74 educators. The original targeted number was 364 educators and no specific target was set for the students. Knowing that the WMS intake was 177 graduates each year, the student respondents represent almost 10% of the students population and the educator respondents represent almost 20% of the educators population. From these percentages, it is noticed that the educators were more cooperative in this study with higher number of responses and shorter response times than the students.

The surveys included questions about the course the students were studying or the course educators were teaching. The majority of the students who participated in this study were students on the Bachelor of Medicine and Surgery (MB

ChB) programme. Educators involved in this study have experience teaching MB ChB courses and postgraduate courses for varying numbers of years ranging from 1 to more than 20 years. The survey answered by the students and educators contained common questions for collecting data about the respondents participating in order to conduct the proper analysis for different sections of the medical education community. These common questions were categorized according to the KAP survey sections.

3.4.1 Knowledge

The focus of this research is to integrate open educational data on the web. Therefore, it was beneficial to know whether the WMS community is aware of the existence of the Open Educational Resources (OER) movement in general.

K1: *Prior to receiving this survey, were you aware of the existence of the open educational resources initiative and the existence of free educational content in Learning Object Repositories?*

Table 3.1: Percentages of students and educators awareness about OER

	Educators	Students
Yes	28%	37%
No	72%	63%

The responses from both the students and the educators are presented in Table 3.1. The majority of the responses about knowledge of OER are negative for both students and educators of WMS, and the majority of both students and educators were not aware of the OER movement. Although the respondents use the open web for learning, they are not aware of the existence of such movements or their purposes. The results are encouraging to proceed with this research as it will

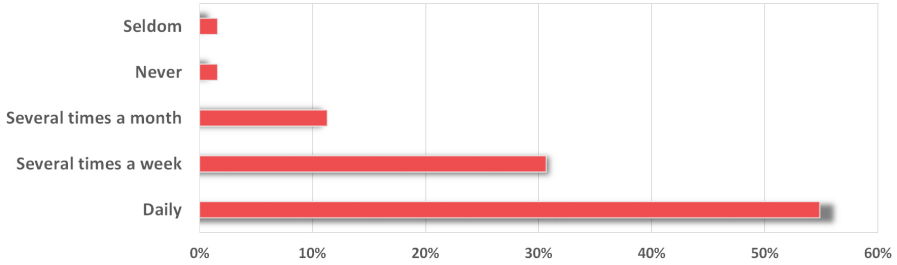
attract more attention to open data on the web.

3.4.2 Practice

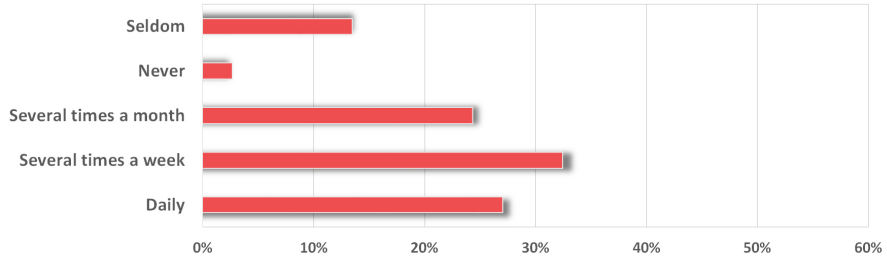
The survey included a set of questions asked to record the practices of students and educators when searching for online content. How often do they search, where, and what type of search criteria do they find beneficial?

P1: *How often do you search for educational material online?*

The results presented in Figure 3.1 illustrate a difference in the search frequency between the students and the educators. Students tend to search consistently for information to help them learn. More than 80% of the student respondents search at least several times a week compared to around 60% of the educators respondents. Another important aspect in the practice part is where the respondents find educational content.



(a) Students



(b) Educators

Figure 3.1: The frequency of searching activities

P2: *Which websites do you currently use to look for educational content?*

It is obvious that both students and educators rely on the popular search engines for discovering content on the web from the percentages of both students and educators shown in Table 3.2. Then, the respondents were asked whether they use the filtering criteria provided in the websites they search.

Table 3.2: The websites used by the respondents for searching

	Educators	Students
Popular search engines, such as Google, Bing, or Yahoo	92%	89%
Others	8%	11%

P3: *Advanced search features in search engines and online libraries allow you to filter the search results retrieved according to specific criteria. Do you use such features for filtering results in your regular internet search activities?*

Comparing the results presented in Table 3.3, students showed less experience with filtering criteria than the educators. The nature of the current filtering criteria might not be practical for users. Hence, respondents were asked in the next section, in the attitude section of the KAP survey, if they would change their usage of filtering criteria if new filtering criteria were suggested.

Table 3.3: Percentages of respondents using filtering criteria

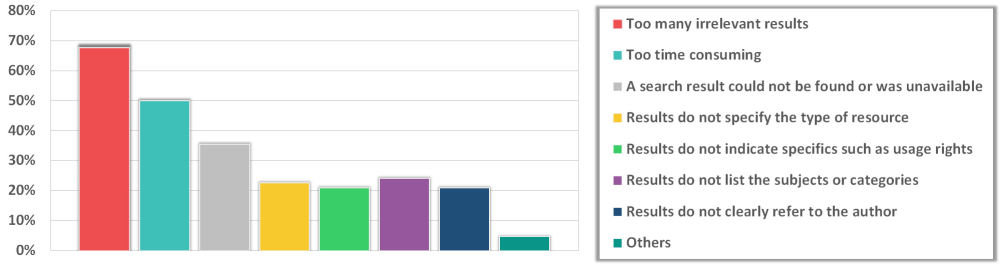
	Educators	Students
Yes	58%	29%
No	42%	71%

3.4.3 Attitude

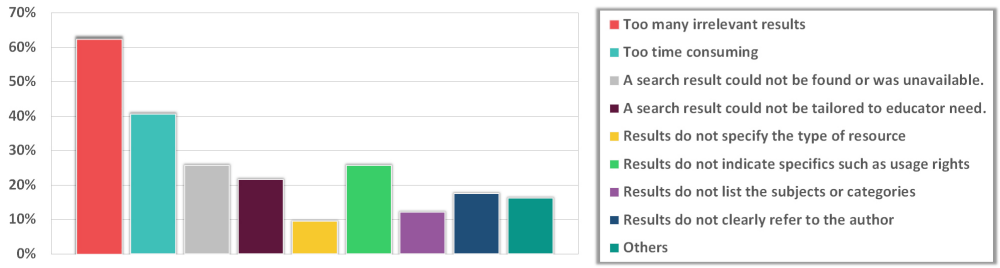
A set of questions in the survey focused on collecting information about the major frustrations and attitudes of students and educators towards their current search habits. The major frustrations for both students and educators were investigated in the first question.

***A1**: What are your major frustrations when searching for educational resources online?*

The results of the major frustrations faced by the respondents are presented in Figure 3.2. Both students and educators have similar rankings for possible frustrations faced while searching. In both sections, "Too many irrelevant results" was considered the main frustration for them when searching for online content. Then the respondents were asked whether some special criteria can be helpful in the search process.



(a) Students



(b) Educators

Figure 3.2: Major frustrations for the respondents when searching

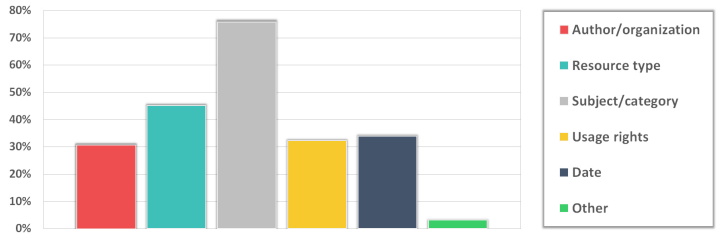
***A2**: If you have more filtering criteria, such as media type, subject area, author name, usage rights, would that improve your search results when searching for online for educational content?*

The results of the students and educator preference to use new criteria such as subject and usage rights are mainly positive responses as illustrated in Table 3.4. The respondents then were asked what criteria they prefer to use if possible.

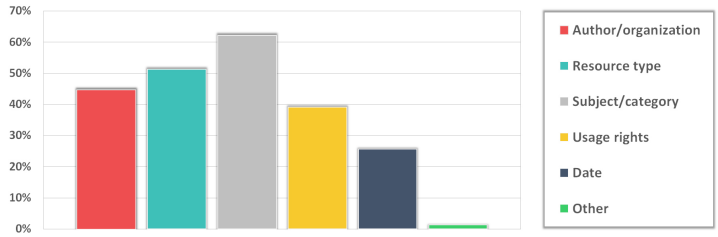
Table 3.4: Percentages of respondents who prefer using new filtering criteria

	Educators	Students
Yes	82%	79%
No	18%	21%

***A3**: What search criteria would you find most helpful in searching for educational resources?*



(a) Students



(b) Educators

Figure 3.3: Preferences of search criteria to be used

The results presented in Figure 3.3 are similar in their order of importance for both students and educators. The subject category is considered the most important filtering criterion for both students and educators.

3.5 Discussion

Although this exploratory study did not cover a large number of the targeted population, the results present a reasonable picture of the current practices and attitudes of students and educators when searching the web for educational content. It appears that both students and educators are not aware of the new trend of open content available in educational institutes. A total of 32% of the respondents are aware of the existence of the OER movement, 37% of the student respondents and only 28% of the educator respondents.

Learning about the respondents' current practices when searching the web deepens the understanding of the respondents' attitudes towards searching. Of the student group, 55% reported that they search the web daily for educational content, and 31% are searching several times a week. In total, 86% of the student respondents search the web at least several times a week for educational content. The educators indicated lower percentages where 27% search the web daily and 32% search several times a week, giving a total of almost 60%. The results are logical since students require more information about a subject to understand it fully. The second important issue in the practices of the respondents is where they search. Almost 90% of the respondents used popular search engines such as *Google*[3], *Yahoo*[4], and *Bing*[5]. This percentage justifies the responses regarding the major frustrations faced when searching the web in the attitude section. Other libraries and websites were reported to be used by both educators and students when searching. The

[3]https://www.google.com/
[4]https://www.yahoo.com/
[5]https://www.bing.com/

PubMed library[6] is one of the most widely mentioned sources when searching for educational content by both educators and students. Also, *Google Scholar* was mentioned when searching for articles on the web with one problem faced, that is not all the results are openly available for public use. Students' responses indicated the use of websites hosting web 2.0 objects when searching for educational content such as *YouTube*[7] and journals websites, in addition to community developed websites such as *Medscape*[8]. Finally, the respondents reported about their current usage of advanced search criteria when searching the web. Educational libraries and general search engines support filtering the search results using advanced criteria to avoid the frustrations resulting from searching the web. The percentage of the student respondents using such filtering criteria is only 29%, compared to a higher percentage of 58% of the educators utilising such advanced criteria to enhance their search process.

The attitude section of the KAP survey collects data about the respondents' feelings and preferences during the search process. When asked about the frustrations they feel when searching the web, students and educators checked all that applies from a list of possible problems web users have. The order of the problems from the most frustrating one, having the highest number of respondents expressing it as a frustration, to the lowest one, is similar for both students and educators group. The problem considered as the major frustration among all the respondents is "Having too many irrelevant results when searching the web for educational content", with percentages of more than 70% in the students group and more than 60% in the educators group. The second and third highest problems expressed as frustrations are that the search process is too time-consuming, and the search process sometimes fails to find what they are searching. From the list of possible frustrations presented in the answer list are answers related to the inability to

[6]http://www.ncbi.nlm.nih.gov/pubmed
[7]https://www.youtube.com/
[8]http://www.medscape.com/

search or filter the results according to specific properties such as the author, the subject, the usage rights, and the type of each single result retrieved. Among those possible frustrations, a low number of respondents indicated that the lack of such properties is considered a frustration for them. Current practices of describing educational content might not include specific properties such as its type or usage rights. Nevertheless, the interesting part from analysing the responses of user frustrations illustrated in Figure 3.2 is that the student group expressed that failing to provide the subject or category property of the search results is the highest frustration of this group of properties. While educators indicated that failing to provide the usage right on the search results is their highest frustration. If specific advanced filtering criteria were presented for students and educators to enhance their search experiences, such as filtering the search results based on its type, subject area, or usage rights, the respondents indicated a very positive attitude towards using such criteria. Both students and educator expressed their interest in using such criteria with percentages equal to 79% of the student group and 82% of the educator group. Proposing a set of possible advanced filtering criteria, the respondents selected all that applied from this list if they thought it would aid them when searching the web.The list of possible advanced search filtering criteria includes the author, the type, the subject, the usage rights, the date, and other criteria that they can specify. The highest criterion preferred by the respondents is the subject criterion with percentages of more than 70% and 60% of the students and educators respectively. The second highest criteria are the type of the resource in the search list retrieved for both groups. The educators' responses order the rest of the criteria as the following: the author, followed by the usage right and finally the date while the students' responses had very close percentage results for these three criteria.

The results collected from the attitude section of the KAP study revealed valuable information about the respondents' problems and preferences in the search process.

3.6 Recommendations

The outcomes of the KAP survey conducted by students and educators of WMS support the research objectives and confirm the gaps detailed in the literature review about current metadata practices in medical educational libraries. Searching the web for content to be used in learning and teaching medicine is the topic being investigated in this survey. Thus, associating the findings of this survey, related to the practices and attitudes of students and educators, with the main problem identified in chapter 1, reinforces the objectives of this research. The most significant results were used to confirm and direct these research activities. The following recommendations pointed the way forward and had an impact on the development of this research.

- *Integrating educational content used for learning from distributed web data sources:* the new practices of concept-based learning (instead of content-based learning) encourage the use of different types of educational materials when teaching or learning a topic. The large use of general purpose search engines such as the websites listed in the survey for searching for educational content indicates that the learners are open to different types of information about the topic searched. Some of the other websites explicitly stated in the survey as being used when searching for educational content are websites hosting only videos, or blogs, or articles written and shared openly with the public and not published in journals or libraries. For example, Google Scholar was mentioned as a source when searching for articles, yet the access rights for the search results is not explicitly mentioned. Thus, integrating educational objects from different distributed open websites into one searchable repository is one of the main objectives of this research.

- *Having easily discoverable educational content on the web*: the primary issue in searching causing major frustrations for web users is what search terms to use

when searching for a particular resource. Hence, the original problem lies in describing the content before publishing it on the web. The lack of providing the right information about the content being published lessens the probability of it being found. The new trend of the web of data aims to publish raw data on the web making it easily accessed and searchable. Although the community being studied in this research is not aware of other related movements such as the OER movement, more projects and campaigns are focusing on increasing awareness about it. Initiatives such as open education were the driving force behind conducting this research in its early stages.

- *Introducing new features for organizing and filtering the search results*: the heavy usage of popular search engines that are not specified for educational purposes is not efficient for students and educators. The time spent on navigating the results and filtering them is wasted and stated as one of the major frustrations when searching. The new criteria proposed in the survey had a significant acceptance from the participants as being beneficial for filtering the search results. This research considered incorporating the most important criteria suggested in the survey. The objectives of this research are to develop and enhance the search filtering criteria based on the community needs and preferences.

3.7 Summary

This chapter has described the exploratory study performed to gather information about the problems and initial recommendations from the domain of interest in this research. The knowledge, attitude, and practices (KAP) of students and educators in the Warwick Medical School has been examined. The KAP survey has been used to extract information from the domain of study to investigate their frustrations and needs when it comes to searching the web for educational content. A list of

recommendations has been formulated from the opinions collected in this study. The recommendation are the main outcomes of this chapter and have been used as guidelines for designing a solution for the problem investigated.

In conclusion, this chapter is concerned with studying a sample form medical education domain in order to address the research objective **O2**: "Identify the search practices and challenges faced by students and educators in the medical education field when searching for educational content on the web including Web 2.0 sites and online academic libraries". By addressing this research objective, this chapter completes the answer of the research question that was partially answered in chapter 2, that is research question **R1**: "What are the current metadata schemas used when publishing medical educational content on the web and what are the essential elements from the user perspective when searching for such content?". The recommendations from this KAP survey have defined the guidelines for developing the solution proposed in following chapters.

Chapter 4

Describing Educational Medical Objects

4.1 Introduction

The advent of the Open Data revolution is undoubtedly changing the way education is delivered. Books, articles, videos, and other types of educational content are increasingly being published freely on the web. The growing content of which is further supported by new learning theories such as connectivism [Siemsens, 2005] confirming that, in this digital age, learning is established through connecting contents to concepts. Nowadays, any digital resource, named learning object [Wiley, 2011], can be used to facilitate learning. Search engines such as *Google*, *Yahoo*, and *Bing* enable access to large amounts of open content, yet they are not efficient for retrieving relevant information for a particular purpose and require significant efforts for users to search, find, and link relevant learning objects of different multimedia types together. Participants in the exploratory study (Chapter 3) expressed their frustrations when searching the web. Educational organisations have different publishing requirements that govern the organisation of its content. Metadata is considered the key component for publishing and managing online content. It has

been defined as "data describing the context, content, and structure of records and their management through time" [Franks and Kunde, 2006]. The term metadata is employed in different domains and is often defined as data about data [Turner, 2002]. In libraries, metadata is used to describe any formal scheme that organises information to describe a resource digital or not digital [Guenther and Radebaugh, 2004]. In the field of e-learning, the metadata standardisation is a highly researched topic, and it has been evident that there is no ideal standard that fits all the organisations' requirements [Devedžic, 2006]. Failing to find a mature metadata standard that satisfies the needs of specific community needs advice the process of developing a new metadata schema, yet it can be a complex and demanding task to master [Diamantopoulos et al., 2011]. Therefore, the metadata community introduced the term Application Profile (AP) that is applied to support the process of tailoring existing metadata schemas for specific applications. General cross-domain metadata standards such as the Dublin Core (DC) metadata [Powell et al., 2007] supports the development of APs extending its elements which enables its wide adoption across different domain [CWA, 2006].

In this chapter, a metadata schema is proposed for describing Educational Medical Objects (EMOs) of different types such as videos, blogs or articles named the Linked Educational Medical Object (LEMO) metadata. It aspires to accommodate describing different types of EMOs and to introduce new elements that can improve the description of the EMOs with semantics using biomedical ontologies. Hence, before starting the design process for a new metadata schema, it is necessary to analyse existing metadata schemas that can be adapted for designing the proposed LEMO metadata [Hodgson, 2008]. Building a new metadata schema requires lots of efforts in comparison of adopting a metadata schema that is well-modelled and supported. Hence, it is beneficial to introduce the concept of Application Profile, as given in the terms definitions section in (ISO 23081)[1]. An Application Profile (AP) is con-

[1]http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=40832

sidered a type of a metadata schema that consist of metadata elements drawn from existing metadata schemas and combined with possible new elements introduced and optimised for a particular application [Heery and Patel, 2000]. The idea from the application profiles is to use existing well-established metadata schemas and to introduce new refinements into its elements that help in achieving the functional requirements of a particular application.

4.1.1 Chapter Objectives

This chapter aims to present the methodology followed for designing and implementing the LEMO metadata schema as an application profile, and it addresses the research objectives **O3** and **O4**. The methodology maps to these research objectives starting with the research objective **O3**: "Conduct a comparative analysis of the existing metadata schemas to identify the common characteristics for describing EMOs" at the early stages of designing the LEMO AP. The later stages followed in the methodology achieve the research objective **O4**, that is concerned with designing the LEMO metadata, implementing it, and validating its usage by experimenting with real data collected from web data sources.

4.1.2 Chapter Outline

This chapter explains the phased development of the LEMO metadata schema and is organised as follows. It begins with a detailed demonstration of the design methodology adopted for proposing the LEMO metadata. Followed by a domain analysis and requirements specification that are necessary for modelling the metadata. The LEMO metadata schema and the techniques for implementing it are explained next. Finally, experimental testing with data collected from candidate web data sources is necessary to validate the LEMO metadata applicability.

4.2 Design Methodology

This research is applied in the medical education field. Hence, the needs of this community are taken into consideration in the design methodology. The LEMO metadata schema was created as a Dublin Core Application Profile (DCAP) following the recommendations presented in the Singapore Framework (SF) for defining DCAPs [Nilsson et al., 2008]. The DC metadata is a simple, flat, and descriptive metadata schema that consists of 15 elements covering the medical community needs. Such features provide the flexibility of tailoring and customising of its elements to satisfy the needs of the medical education community. The Singapore Framework defines the components that are used for documenting the process of developing DCAPs. The framework describes a DCAP as a packet of documentation containing the following mandatory parts and two more optional components that are usage guidelines and encoding syntax guidelines [Nilsson et al., 2008].

- *Functional requirements:* it describes the needs of developing the DCAP. The detailed application requirements or the community needs are listed. This part should also demonstrate the missing features that the AP is supposed to cover.

- *Domain model:* it is a formal approach to define the entities to be described and their relations. It is considered a basic blueprint for the developing the application profile.

- *Description set profile:* it describes the AP element set based on the DC element set. It defines the constraints on what attributes to use and how their values may be referenced.

The detailed steps needed for developing the LEMO metadata schema based on the Singapore Framework for developing any DCAP are detailed in Table 4.1.

These phases document the process of building the LEMO metadata schema as a DCAP and follows a top-down approach.

Table 4.1: Phases and tasks of designing the LEMO metadata schema as DCAP

Phase 1. Domain analysis and requirements specification
• Summarize the needs of the medical education community based on the recommendations of the exploratory study.
• Conduct a comparative study of the current practices applied for organising medical web data sources.
• Relate the needs of the medical education community to the common attributes discovered from the comparative study.
• Produce a list of the functional requirements of the LEMO metadata schema.
Phase 2. Metadata design and modelling
• Map the functional requirements defined to the DC metadata elements.
• Decide which elements to use from the DC metadata element set.
• Devise a domain model introducing the newly proposed elements and its relations to the DC elements.
• Elaborate on the usage of each element in the LEMO metadata schema.
• Produce a domain model that defines the entities to be described and their relations.
Phase 3. Developing and testing the metadata
• Develop the RDF/XML format that defines the domain model designed for the LEMO metadata schema.
• Develop the needed techniques for mapping the EMOs collected from the web into the LEMO metadata schema.
• Collect examples of possible educational medical objects from web data sources to test the LEMO metadata schema.
• Conduct experiments to test the validity of mapping and representing different metadata schemas using the LEMO RDF/XML metadata schema.

The ensuing sections of this chapter describe the SF-based LEMO metadata schema development process and present the details related to requirements elicitation, metadata design, and metadata development tasks.

4.3 Domain Analysis and Requirements Specification

In order to specify the functional requirements of the LEMO metadata, an analysis of the application domain was necessary. The analysis involves studying the needs of learners in the medical education community, and analysing existing metadata schema applied in that field. The Warwick Medical School (WMS) students and educators participated in the exploratory study conducted earlier in this research (chapter 3). The needs and frustrations discovered by this exploratory study present the guidelines for identifying the functional requirements of developing the DCAP. Additionally, existing metadata schemas that are applied to existing online medical libraries have been analysed and compared to support the process of requirements identification. This section details the needs discovered from the exploratory study conducted with the WMS community. It also presents a summarized comparative analysis of existing metadata schemas in practice. Finally, the functional requirements for the LEMO metadata schema are elicited based on the results of the domain analysis.

4.3.1 Medical Education Domain Analysis

In order to identify the needs of the medical education community, the recommendations resulted from the WMS exploratory study (chapter 3) has been considered. The exploratory study revealed valuable information about the practices and attitudes of the medical education community regarding current and suggested techniques for searching EMOs on the web from which the following needs has been deduced. The following list has guided the process of eliciting the functional re-

quirements for the proposed LEMO metadata schema.

- Integrating distributed web data sources: the participants in the exploratory study had stated the use search engines such as *Google* to find information in addition to other specialised educational libraries such as *PubMed Library*. Hence, the need to integrate possible EMOs from various web data sources is evident. The integration process requires having a flexible metadata schema that can accommodate describing different types of EMOs such as videos, blogs, and articles.

- Enhancing the discoverability of the EMOs: the medical education community had a heavy emphasis on the importance of enhancing the search and organisation of EMOs based on its subject. Hence, the subject attribute used to describe the EMOs is one of the important features to consider in the LEMO metadata schema in order to provide better discoverability when searched.

- Introducing new attributes that improve organising the EMOs and filtering the search process: the results of the exploratory study had revealed valuable information about the participants' attitudes towards using filtering criteria. Filtering criteria such as usage rights, type of EMOs, or the subject it belongs to are among the most important attributes considered in the exploratory study. Introducing such attributes in the LEMO metadata schema enables using these features for filtering the search results.

4.3.2 Analysis of Current Practices

A comparative study is conducted in order to identify common elements used in the different APs developed in the field of medical education. Four metadata schemas have been detailed in the background research (chapter 2) and were considered as part of the comparative analysis conducted in this phase of developing the LEMO metadata schema. These schemas were applied in educational libraries in the med-

Table 4.2: Results of comparative analysis of medical metadata schemas

	HealthCareLOM		mEducator		HEAL		NLM	
1	General	PN	General	PN	General	PN	General	-
2	Identifier	LN	Identifier	LN	Resource URN	LN	Identifier	PN
3	Title	LN	Title	LN	Title	LN	Title	SN
4	Description	LN	Description	LN	Description	PN	Description	PN
5	Lifecycle	PN	Lifecycle	PN	Creation Date	LN	Date	PN
6	Rights	PN	Rights	PN	IPR	LN	Rights	SN
7	Resource Type	LN	Resource Type	LN	Resource Type	LN	Resource Type	SN
8	Keywords	LN	Keywords	LN	Keywords	LN	Keywords	LN
9	Classification	PN	Classification	PN	Classification	PN	Subject	PN
10	Educational	PN	Educational	PN	Educational	PN	Educational	-
11	Relation	PN	Relation	-	Companion	PN	Relation	SN
12	Technical	PN	Technical	PN	Technical	LN	Technical	-

ical education field. The four metadata schemas are HealthCare LOM, mEducator, HEAL, and NLM. These schemas are developed as Application Profiles of existing well-established metadata. More details about the aims of developing these APs and the organisations responsible for creating and maintaining them have been explained in the background research (Chapter 2) presented in the thesis. The analysis of these four schemas compared the core elements composing each metadata schema. Each metadata schema consists of a set of elements and rules that govern their relations where an element might be represented as a Parent Node (PN), Leaf Node (LF), or Single Node (SN). The definition of each node is as follows. A Parent Node (PN) is an element that is composed of a set of elements that are leaf nodes in the schema. A Leaf Node (LF) is an element that is a child node of another PN and has no children, while a Single Node (SN) is an element that is descendent directly from the root and is not composed of any children. The comparative analysis results are outlined in Table 4.2 where the core element of each AP are elicited along with how they are represented in that AP.

The analysis result has shown that the four metadata schemas have common elements between them. Identifier, Title, and Description are core elements in all of the four schemas. Mostly, they are represented as LN or PN. Additional common elements are Date, Intellectual Property Rights (IPR), Type, Keywords, and Clas-

sification elements. The analysis revealed that some elements such as Educational, Relation, and Technical are not implemented in all four schemas as shown in Table 4.2. Such elements indicate that the requirements of the organisation or the application of which the AP is developed for plays a vital role in the elements composing its schema. For example, HealthCareLOM AP focuses on providing information about the educational aspect of the object it describes. Meanwhile, NLM AP focuses on providing information about the relation between the objects it describes and does not provide any metadata about its educational aspect.

4.3.3 Functional Requirements Specifications

The domain analysis has been conducted as two parts: studying the attitudes and preferences of the medical education community, and the analysis of existing practices in the field of medical education. Based on the facts discovered from the domain analysis, the functional requirements have been specified for developing the LEMO metadata schema. The task of eliciting the functional requirements was based on use cases formulated from the domain analysis. These use cases are detailed in Table 4.3 and form the functional requirements in the LEMO AP development process. The requirements are useful for modelling the LEMO metadata schema, regarding identifying the necessary elements in it or the need for extending some elements to achieve the requirements of the AP being developed. Specifying the requirements and documenting the domain analysis results strengthened the commitment towards developing a full DCAP, offering the shift from the DC flat metadata schema to a more complex one that provides extensible and semantically richer metadata elements.

Table 4.3: Use cases representing the functional requirements of the LEMO AP

	Use case description
1	Facilitate the description of EMOs collected from various web data sources of different types.
2	Enable automatic harvesting and mapping of metadata from web data sources into the LEMO metadata schema.
3	Enable semantic enrichment of metadata elements that will aid in enhancing the description of EMOs and improve its retrieval.
4	Support the categorization of EMOs based on its subjects, type, or usage rights.
5	Support building connections between EMOs that were not previously linked.

4.4 Metadata Modelling

Based on the functional requirements specified in the previous section, the LEMO metadata element set was identified. It builds on the DC element set and extend some of its elements. New features were introduced to the flat structure of the DC metadata to enable a richer description of the EMOs metadata. Figure 4.1 illustrates the domain model of the metadata. It identifies what are the entities to be described in the metadata and the relationships between these entities. It is considered as a blueprint for the development of the LEMO AP that defines what

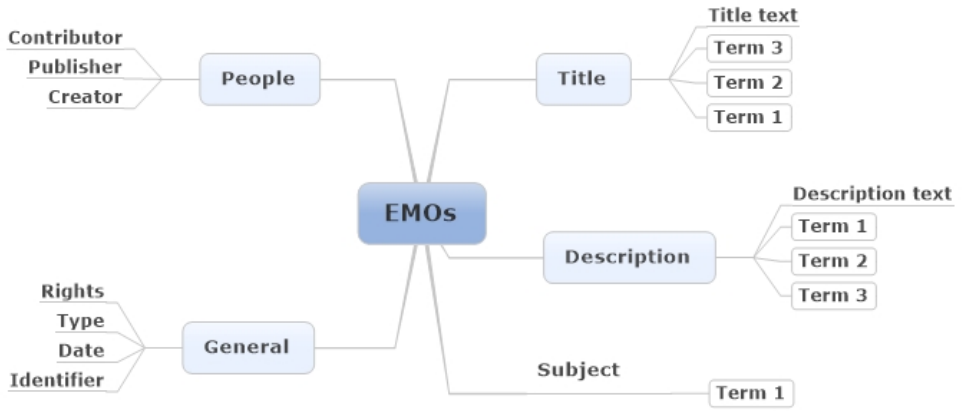


Figure 4.1: LEMO metadata schema conceptual model

DC metadata elements to use or extend. One metadata record of an EMO need to describe the following:

- **General:** attributes that include an identifier, Date, Type, and Usage rights of the EMO.
- **People:** it describes people involved in creating the EMO such as the authors, publishers or contributors.
- **Title:** it provides the title of the EMO and possible keyword terms discovered in the title.
- **Description:** it provides the description text of the EMO and possible keyword terms discovered in that text.
- **Subject:** it used to reference some of the keyword terms discovered in the title or the description that might represent the subject of the EMO.

The LEMO AP functional requirements focused on enhancing the description of EMOs with added semantics that can aid its categorizing and linking. Therefore, the domain model emphasised the need to describe the title and description of EMOs as entities that need to be enriched with further information that can be used for categorizing the EMOs and interlink them. Implementing the LEMO AP is based on exposing the metadata in linked data practice that enables easier integration with external data on the web. The next section explains how the LEMO AP is developed. It details the metadata elements used for describing the entities identified in the domain model above. It identifies the DC metadata elements that have been used in the LEMO AP element set and introduces new elements for extending some of the DC elements.

4.5 Metadata Implementation

The standard foundation for developing any DCAP is Resource Description Framework (RDF). RDF is a language for describing resources published on the World Wide Web [Yu, 2011]. It is a standard published by the W3C[2] in 1999. The RDF standard was updated after introducing the concept of the semantic web in 2001. It supports the vision of the semantic web that is to make the web machine understandable. RDF to the semantic web is similar to what HTML has been to web [Yu, 2011]. Detailed explanation about the RDF standard and how it works has been presented in the background chapter (chapter 2) as part of the web publishing section.

The process of developing any DCAP involves incorporating any terms that are defined using RDF. It can combine terms from multiple namespaces as needed [Coyle and Baker, 2009]. One of the main reason for developing a DCAP instead of creating a new metadata schema is maintaining the interoperability of metadata description across the web. Utilising existing namespaces ensures the interoperability of the metadata records. The LEMO AP was created by taking elements from the RDF namespace[3] and the DCMI namespace[4]. Additionally, a new namespace was established to group the attributes for the purpose of developing the LEMO AP named LEMO namespace.

The LEMO AP element set is illustrated in Figure 4.2. The prefixes **(rdf)**, **(dc)**, and **(lemo)** represent the RDF, DCMI and LEMO namespaces respectively. Figure 4.2 represent the metadata elements describing an EMO in RDF. The figure can be translated into RDF statements where the EMO resource is the subject of the statement identified by a URI. The subject of the statement is described using predicates detailed on the lines connecting the figure elements. The predicates are terms that are also identified by URI using the prefixes of the namespaces they

[2]http://www.w3.org
[3]http://www.w3.org/1999/02/22-rdf-syntax-ns
[4]http://purl.org/dc/elements/1.1/

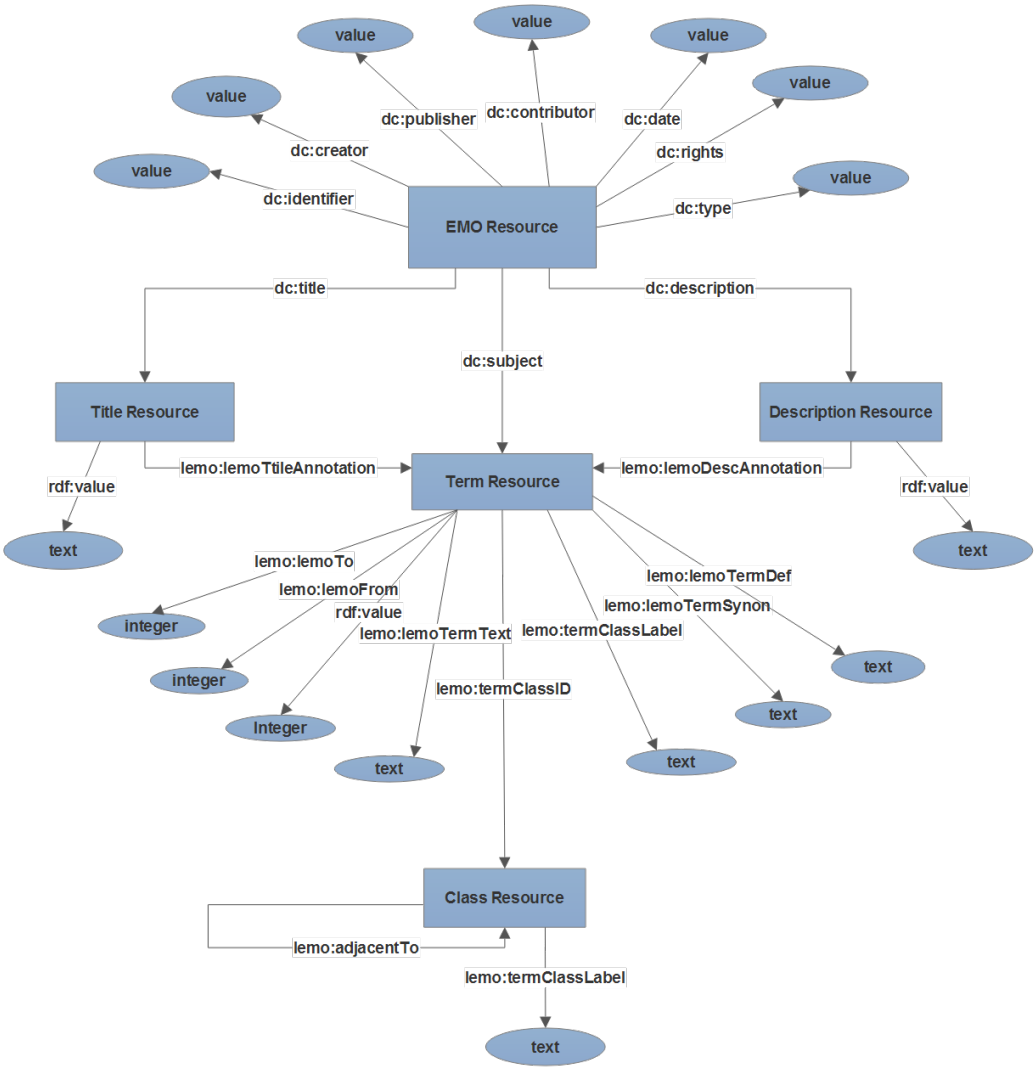


Figure 4.2: LEMO Application Profile (AP) element set

represent. Finally, the object in the RDF statement represent an entity that is connected to the subject resource (EMO) via the predicates. Objects can be literal values or RDF resources represented in URIs. In Figure 4.2, an RDF resource is represented as a rectangle, the predicates are the labels on the lines, and the objects can be either an RDF resources or a literal value represented in the oval shape. The definition of the core entities, which are RDF resources in the figure, and the main metadata elements, represented as predicates, are defined as follows.

- *The EMO:* constitutes the core entity described using the LEMO metadata schema. The EMO is a resource that is described using the DC metadata elements: Identifier, Title, Description, Date, Rights, Type, Author, Contributor, Publisher, and Subject. The values of these elements differ and are explained in the rest of the list.

- *Identifier:* is a DC element that is used to identify the EMOs with distinct URI for each EMO.

- *Creator:* is a DC element that stores the creator of the EMO whether it is a person name or an organisation.

- *Contributor:* is a DC element used to store additional information about the contributors in making the EMO.

- *Publisher:* is a DC element stores the organisation or person name responsible for publishing the EMO.

- *The title:* is an element composed of the title text and further annotations of that text represented in a collection of term resources.

- *The description:* is similar to the title element. It describes the text of the description and the annotations made to that text represented as term resources.

- *The term:* is a newly introduced resource used to provided a detailed description of the text annotated in the title or the description elements related to it. Also, it provides details about the concept in the ontology annotated by that term.

- *The class:* is a newly introduced resource that stores the relations between the concepts of the ontology as it is described in the ontology hierarchical relations.

- *Rights:* is a DC element used to store the usage right of the EMO.

- *Date:* is a DC element record the dates of creating or revising the EMO it describes.
- *type:* is a DC element stores the type of the EMO.
- *Subject:* is a DC element used to represent the terms that are selected as the categories classifying the EMO.

The newly introduced metadata elements, which are predicates of the (lemo) namespace, are used to enrich the EMOs title and description elements that are represented as RDF resources instead of literal values. The LEMO focused on extending the DC metadata elements as shown at the bottom part of Figure 4.2 to use concepts from external biomedical ontologies to annotate and link the literal values in the LEMO metadata with ontology classes. The following sections explain the RDF/XML format used to represented the LEMO AP and the mapping process developed for mapping different XML formats into the RDF/XML format of the LEMO AP. The EMOs collected from the web are initially encoded in XML. Hence, the process of describing them using the proposed LEMO AP requires mapping the metadata from one format to another.

4.5.1 RDF/XML metadata

This section explains the RDF statements used to describe an EMO based on the LEMO metadata schema. In this research, an RDF model stored all RDF statements that describes EMOs metadata. At start, the RDF/XML sample provided in listing 4.1 indicates the terminology used to explain the RDF statements that describe an EMO based on the LEMO metadata schema. Statements in RDF are used to describe a resource in the collection of EMOs harvested from the web, with the resource being the subject of the statement. In the RDF/XML listing 4.1, the `rdf:RDF` statement (line 1) indicates that this XML file is intended to represent an RDF model and it ends in line 7 with the end tag `</rdf:RDF>`. Line 2 represents the

XML namespace declaration via the `xmlns` attribute which specifies that the prefix `rdf:` is used to represent the predicates' URIs if the predicate is tagged with this prefix. Now the following RDF statements (line 3-6) describe a resource as being the subject of the statement. The term `rdf:Description` indicates the beginning of a description of a resources, and the attribute `rdf:about` is used to determine the URI of that subject resource as shown in line 3. The URI can be a URL or an invented unique URI added to the RDF model. The tag `</rdf:Description>` indicates the end of the resource description as shown in line 6. Now the statements in between are used to describe the predicates and the predicates' values of the statement. Line 4 describes one predicate that has another resource as its value. The value of a predicate is referred to as the object of the statement. The predicate is represented by a URI using the prefixes list provided at the beginning of the RDF model. Another predicate example is shown in line 5 and its value is a literal value not a resource.

Listing 4.1: RDF statements example

```
1 <rdf:RDF
2     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3 <rdf:Description rdf:about="URI of the statement's subject">
4     <predicateURI rdf:resource="URI of the statement's object" />
5     <anotherPreducateURI>Literal value</anotherPredicateURI>
6 </rdf:Description>
7 </rdf:RDF>
```

Now that the terminology and the structure of an RDF statement has been clarified, the RDF statements used to describe the resources represented in th LEMO AP element set in Figure 4.2 are explained.

The main resource composing the LEMO AP is the EMO resources. The full description of an EMO resource in RDF/XML syntax is detailed in listing 4.2. The listing represents the RDF statements that translate part of the graph concerning the EMO resources and its related predicates. The EMO resource is described using DCMI element set represented as predicates with the DC prefix. Some predicates can

have multiple related objects such as the <dc:creator> term. The objects in most of the RDF statements describing the EMO resource are literal values that can be either text or URLs, except for the <dc:title> (line 11) and <dc:description> (line 12) predicates. These two predicates are related to objects of type RDF resource that are identified by unique URIs. Also, the object related to the <dc:subject> predicate (line 16) is an RDF resource that represent a term related to the title or description resource. The detailed description of each RDF resource and its related predicates are be explained shortly.

Listing 4.2: EMO metadata schema

```
1  <rdf:RDF
2      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3      xmlns:dc="http://purl.org/dc/elements/1.1/"
4      xmlns:lemo="http://www.warwick.ac.uk/ias/lemo/">
5    <rdf:Description rdf:about="resourceURI">
6      <dc:identifier> http://www.somewhere.com/uri </dc:identifier>
7      <dc:creator> First Author </dc:creator>
8      <dc:creator> Second Author </dc:creator>
9      <dc:contributor> Name </dc:contributor>
10     <dc:publisher> Name </dc:publisher>
11     <dc:title rdf:resource="resourceURI:title"/>
12     <dc:description rdf:resource="resourceURI:desc"/>
13     <dc:date> 9-Nov-2015 </dc:date>
14     <dc:rights> Free </dc:rights>
15     <dc:type> Article </dc:type>
16     <dc:subject rdf:resource="resourceURI:title:term"/>
17     <dc:subject rdf:resource="resourceURI:desc:term"/>
18   </rdf:Description>
19   </rdf:RDF>
```

The RDF/XML representation of the resources acting as objects for the <dc:title> and the <dc:description> predicates are detailed in listings 4.3 and listing 4.4 respectively. The two resources are similar in their functionality since they both provide the literal value of the title or the description, and introduces new attributes that are used to store annotations discovered in the literal text using external biomedical ontologies.

The Title resource is given a distinct URI derived from the EMO URI it describes and appended with the text *":title"* as illustrated in the listing 4.3. This resource is described by two predicates , the `<rdf:value>` that stores the literal text representing the title of an EMO and the `<lemo:lemoTitleAnnotation>` that is used to relate the Title resource to terms extracted from biomedical ontologies. The terms are discovered from annotating the text of the title with concepts from the ontology. The terms are represented as RDF resources and can be related to the title and the description resources as objects for their `<lemo:lemoTitleAnnotation>` or `<lemo:lemoDescAnnotation>` predicates. The Term resources are explained later. Any Title resource can have more than one annotation discovered. Hence, it can have many predicates relating to many Term resources.

Listing 4.3: Title resource schema

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:lemo="http://www.warwick.ac.uk/ias/lemo/" >
  <rdf:Description rdf:about="resourceURI:title">
    <rdf:value>Title text</rdf:value>
    <lemo:lemoTitleAnnotation rdf:resource="resourceURI:title:term1"/>
    <lemo:lemoTitleAnnotation rdf:resource="resourceURI:title:term2"/>
  </rdf:Description>
  </rdf:RDF>
```

The RDF/XML representation of the Description resource detailed in listing 4.4 is similar to the title resource. The only difference is the name of the predicate used to describe the annotations discovered in the description text that is `<lemo:lemoDescAnnotation>`. The use of different metadata elements for representing the annotations discovered in the title or the description text is necessary for further processing of the metadata. The use of two different predicates points out which of the terms is annotated in the title and which are annotated in the description of an EMO. The number of predicates related to the Description resource is not limited to a particular number as the case in the Title resource.

Listing 4.4: Description resource schema

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:lemo="http://www.warwick.ac.uk/ias/lemo/" >
  <rdf:Description rdf:about="resourceURI:desc">
    <rdf:value>Description text</rdf:value>
    <lemo:lemoDescAnnotation rdf:resource="resourceURI:desc:term1"/>
    <lemo:lemoDescAnnotation rdf:resource="resourceURI:desc:term2"/>
  </rdf:Description>
  </rdf:RDF>
```

The Term resource is used to describe the annotations discovered in the title and the description of an EMO. The RDF/XML representation of the Term resource is described in listing 4.5. The annotations are made to literal text based on biomedical ontologies that are used to enrich the semantic meaning of the text. Each Term resource is given a unique URI derived from the Title or the Description resource it describes as shown in listing 4.5 (line 5).

Listing 4.5: Term resource schema

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:lemo="http://www.warwick.ac.uk/ias/lemo/" >
  <rdf:Description rdf:about="resourceURI:desc:term1">
    <lemo:lemoFrom>9</lemo:lemoFrom>
    <lemo:lemoTo>12</lemo:lemoTo>
    <lemo:LemoTermText>Text</lemo:LemoTermText>
    <rdf:value>1</rdf:value>
    <lemo:lemoTermID>http://ontology.com/classID</lemo:lemoTermID>
    <lemo:lemoTermClassID>http://ontology.com/classID </lemo:lemoTermClassID>
    <lemo:lemoTermClassLabel>Annotation Label </lemo:lemoTermClassLabel>
    <lemo:lemoTermDef>The definition</lemo:lemoTermDef>
    <lemo:lemoTermSynon>Synonums</lemo:lemoTermSynon>
  </rdf:Description>
  </rdf:RDF>
```

The Term resource is described using statements that detail information about parts of the text annotated in the title or the description of an EMO, and information about the ontology class annotating that part of the text. The predicates

related to the Term resources are <lemo:lemoFrom> (line 6) and <lemo:lemoTo> (line 7) that record the indices of the text annotated in the literal text of the resource it relates to. While the predicate <lemo:lemoTermText> (line 8) related to literal value that represent the word annotated in the text using the given indices in the previous predicates. The <lemo:lemoTermID> which gives a unique ID for the term, and <rdf:value> predicate that is used to store a weight value for the Term resource after processing the collection of terms annotating a specific text.

The rest of the predicates are used to describe the information about the ontology class used for annotating that exact text specified. The objects of these predicates are all retrieved from the ontology used for enrichment. The predicate <lemo:lemoTermClassID> stores the ID of the class as it is in the ontology and the <lemo:lemoTermClassLabel> stores the label used to describe that class in the ontology. The last two predicates <lemo:lemoTermDef> and <lemo:lemoTermSynon> stores the definition and synonyms stored for that class in the ontology if exist. The purpose of using the predicates <lemo:lemoTermID> and <lemo:lemoTermClassID> might be confusing to differentiate at this stage of the research. However, possible updates of the LEMO AP includes using multiple ontologies for enriching its records. Hence, the one term represented by the <lemo:lemoTermID> can have several ontology classes annotating it and then having this two predicate will be useful.

The final component of the LEMO AP is the Class resource represented in the RDF/XML listing 4.6. The ontology classes used for annotating the text are parts of an ontology that are hierarchically related. The Class resource is used to describe these relations for further usage in the processing of the terms collections annotated for each EMO. The predicate <lemo:adjacentTo> (line 7) describe the adjacency of one class in the ontology to the rest of the classes annotating the EMOs.

Listing 4.6: Class resource schema

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:lemo="http://www.warwick.ac.uk/ias/lemo/" >
<rdf:Description rdf:about="http://ontology.com/classID">
    <lemo:lemoTermClassLabel>Class Label</lemo:lemoTermClassLabel>
    <lemo:adjacentTo>http://ontology.com/classID2</lemo:adjacentTo>
</rdf:Description>
</rdf:RDF>
```

4.5.2 Mapping process

It was mentioned earlier in the introduction of this section that the EMOs collected from the web are initially encoded in XML. The EMOs are described based on the standards followed by the website responsible for publishing it. Therefore, the aggregated EMOs have heterogeneous metadata formats that need to be mapped into the proposed LEMO AP. The mapping process guarantees that all the EMOs aggregated from distributed web data stores are represented using the LEMO AP implemented in RDF/XML format. Hence, the process of transforming the EMOs metadata collected, from one format to the LEMO RDF/XML format, requires developing a mapping process. This process was developed using the Extensible Stylesheet Language Transformation (XSLT)[5] language that is used for transforming XML documents into other XML formatted documents.

The mapping process has been developed as part of LEMO AP implementation. In this section, the mapping process is explained using a scenario-specific case for a better explanation of the transformations happening. The scenario explains the technique developed, based on XSLT language, for transforming the XML file into the desired LEMO AP format, and it details the final result of the transformation that illustrates the same EMO metadata described using the LEMO AP instead of its original metadata XML records.

[5]http://www.w3schools.com/xsl/xsl_intro.asp

The EMO presented in this scenario case has been harvested from *PubMed Library*. This library provides an interface for harvesting its content using the OAI-PMH protocol [Lagoze and Van de Sompel, 2003]. The metadata records describing this EMO are illustrated in listing 4.7. As shown in the XML listing, the original metadata schema adopted by *PubMed Library* is based on the DC metadata schema.

Listing 4.7: OAI-PMH harvested record

```
<record><header>
  <identifier>oai:pubmedcentral.nih.gov:4114209 </identifier>
  <datestamp>2014-08-12</datestamp>
  <setSpec>frontpubhealth</setSpec><setSpec>pmc-open</setSpec></header>
  <metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" >
  <dc:title>Delay in Breast Cancer: Implications for Stage at Diagnosis and
  Survival</dc:title>
  <dc:creator>Caplan, Lee</dc:creator>
  <dc:subject>Public Health</dc:subject>
  <dc:description>Breast cancer continues to be a disease with tremendous public
  health significance....</dc:description>
  <dc:publisher>Frontiers Media S.A.</dc:publisher>
  <dc:date>2014-07-29</dc:date>
  <dc:identifier>http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4114209/
  </dc:identifier>
  <dc:identifier>http://www.ncbi.nlm.nih.gov/pubmed/25121080 </dc:identifier>
  <dc:identifier>http://dx.doi.org/10.3389/fpubh.2014.00087 </dc:identifier>
  <dc:type>Text</dc:type>
  <dc:language>en</dc:language>
  <dc:rights>Copyright  2014 Caplan.</dc:rights>
  <dc:rights>http://creativecommons.org/licenses/by/3.0/</dc:rights>
  <dc:rights>This is an open-access article distributed under the terms of the
  Creative Commons Attribution License (CC BY). </dc:rights>
</oai_dc:dc>
</metadata></record>
```

The mapping process was developed using Java programming language where XSLT templates have been incorporated in the process of transforming the XML files. The final result of the mapping process for the EMO metadata into the LEMO AP is detailed in listing 4.9 at the end of this section. XSLT templates applied in the mapping process convert the metadata elements from the original XML file into

their corresponding metadata elements in the LEMO AP format.

As an example of the mapping process, converting the <dc:title> metadata element in listing 4.7 (line 8) to its corresponding description <dc:title> in the LEMO AP as represented in listing 4.9 (line 14 and lines 23-25) is explained. The part of the XSLT template responsible for mapping this metadata element is detailed in listing 4.8. The XSLT converts the metadata element and creates a new resource (Title resource) that is related to the title element <dc:title> as its value as shown in listing 4.9 (line 14).

Listing 4.8: XSLT for converting the title attribute

```
1 <xsl:template match="oai:OAI-PMH/oai:ListRecords/oai:record">
2       <rdf:Description>
3       <xsl:attribute name="rdf:about">
4               <xsl:value-of
                select="concat(./oai:header/oai:identifier,':title')"/>
                </xsl:attribute>
5               <rdf:value> <xsl:value-of
                select="oai:metadata/oai_dc:dc/dc:title"/> </rdf:value>
6       </rdf:Description>
7 </xsl:template>
```

Similar XSLT template parts map any XML metadata file into the LEMO AP format. The example above described the XSLT part responsible for transforming one metadata element in the LEMO AP. This XSLT part is part of a larger template responsible for transforming all metadata elements represented in some XML format into the LEMO AP format. The final LEMO AP metadata describing an EMO are listed in listing 4.9. It consists of 3 resources; the EMO resource itself that is described starting from line 7, the title resource of the EMO starting at line 23, and the description resource of the EMO starting at line 27. Reflecting LEMO data model illustrated in Figure 4.2 helps to understand the RDF/XML listing describing an EMO. The rest of the resources and attributes shown in the LEMO metadata schema in figure 4.2 are added after processing the metadata records for enriching and integrating its content using biomedical ontologies. The techniques developed

for harvesting and enriching the metadata of LEMO AP are detailed as part of developing the LEMO system presented in chapter 5.

Listing 4.9: EMO resource described in LEMO metadata schema

```
<rdf:RDF
        xmlns:lemo="http://www.warwick.ac.uk/ias/lemo"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:oai="http://www.openarchives.org/OAI/2.0/"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
        <rdf:Description rdf:about="oai:pubmedcentral.nih.gov:4114209">
                <dc:identifier>http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4114209/
                </dc:identifier>
                <dc:identifier>http://www.ncbi.nlm.nih.gov/pubmed/25121080
                </dc:identifier>
                <dc:identifier>http://dx.doi.org/10.3389/fpubh.2014.00087
                </dc:identifier>
                <dc:creator>Caplan, Lee</dc:creator>
                <dc:publisher>Frontiers Media S.A.</dc:publisher>
                <dc:date> 2014-07-29 </date>
                <dc:title rdf:resource="oai:pubmedcentral.nih.gov:4114209:title"/>
                <dc:description
                rdf:resource="oai:pubmedcentral.nih.gov:4114209:desc"/>
                <dc:rights>Copyright  2014 Caplan.</dc:rights>
                <dc:rights>http://creativecommons.org/licenses/by/3.0/</dc:rights>
                <dc:rights>This is an open-access article distributed under the
                terms of the Creative Commons Attribution License (CC BY).
                </dc:rights>
                <dc:type> Article </dc:type>
                <dc:subject>Public Health</dc:subject>
        </rdf:Description>

        <rdf:Description rdf:about="oai:pubmedcentral.nih.gov:4114209:title">
                <rdf:value>Delay in Breast Cancer: Implications for Stage at
                Diagnosis and Survival</rdf:value>
        </rdf:Description>

        <rdf:Description rdf:about="oai:pubmedcentral.nih.gov:4114209:desc">
                <rdf:value>Breast cancer continues to be a disease with
                tremendous public health significance....</rdf:value>
        </rdf:Description>
</rdf:RDF>
```

4.6 Experimental Testing

One of the functional requirements specified when developing the LEMO AP (section 4.3.3) is describing EMOs collected from distributed web data sources. The LEMO AP should accommodate different types of EMOs harvested from online educational libraries and Web 2.0 data sources. Experiments were conducted to test the LEMO AP ability to handle describing videos, blogs, in addition to articles.

4.6.1 Videos

One type of EMOs that can be used for teaching and learning in medical education is videos. Several *YouTube* channels are managed by educational organisations, and they use them for publishing teaching videos and tutorials. The content of a *YouTube* channel can be harvested via RSS feeds. Syndicating the content of *YouTube* channels can store feeds as XML documents which includes full text and metadata of the posts published by that channel. For example, the video[6] shown in Figure 4.3 can be syndicated into an RSS feeds reader. A sample of the RSS feeds is detailed in listing 4.10 that represents a snippet from the XML document describing the metadata of the video given in the example.

Listing 4.10: RSS feeds for a YouTube video

```
<item>
    <guid isPermaLink='false'>
        http://gdata.youtube.com/feeds/api/videos/aVz-Ja9Grvg</guid>
    <pubDate>Fri, 17 Oct 2014 03:52:32 +0000</pubDate>
    <atom:updated>2014-12-15T03:45:16.000Z</atom:updated>
    <category domain='http://schemas.google.com/g/2005#kind'>
        http://gdata.youtube.com/schemas/2007#video</category>
    <category domain='http://gdata.youtube.com/schemas/2007/categories.cat'>
        Education</category>
    <title>The Diabetic Foot Exam</title>
    <description>An overview and demonstration of the diabetic foot exam,
    including inspection for common deformities, evaluation of vascular supply,
    and screening for neurop...</description>
    <link>http://www.youtube.com/watch?v=aVz-Ja9Grvg</link>
```

[6] https://www.youtube.com/watch?v=aVz-Ja9Grvg

Figure 4.3: YouTube video example

```
13      <author>Erics Medical Lectures </author>
14      <media:category label='Education'
        scheme='http://gdata.youtube.com/schemas/2007/categories.cat'>Education
15      </media:category>
16  </item>
```

4.6.2 Blogs

Blogs are published by medical journal or academic organisations to present case studies, articles, or experimental results such as the blog maintained by the New England Journal of Medicine (NEJM). An example of a case study presented in NEJM[7] is illustrated in Figure 4.4. Using RSS feeds to syndicate the content of this blog, XML documents can be retrieved and stored for further use. A sample of the XML file describing the blog article shown in Figure 4.4 is detailed in listing 4.11.

Listing 4.11: RSS feeds for a NEJM blog

[7] http://www.nejm.org/doi/full/10.1056/NEJMicm1212346

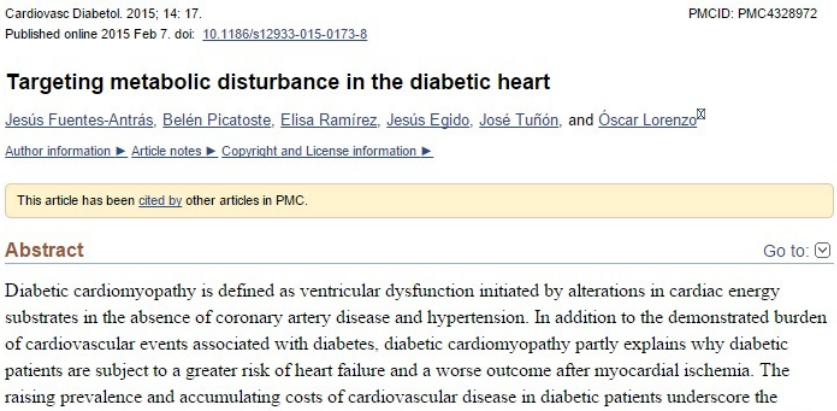
Figure 4.4: Blog example

```
<item rdf:about="http://www.nejm.org/doi/full/10.1056/NEJMicm1212346">
  <title>Pain in the Thumb Related to Disease in the Lung</title>
  <link>http://www.nejm.org/doi/full/10.1056/NEJMicm1212346</link>
  <description>A 59-year-old man presented with a 6-week history of pain in the
  thumb. He also reported having a cough, weight loss, and a history of heavy
  smoking. Physical examination revealed swelling, redness...</description>
  <dc:creator></dc:creator>
  <dc:date>2013-05-02</dc:date>
  <dc:title>Pain in the Thumb Related to Disease in the Lung</dc:title>
</item>
```

4.6.3 Articles

One of the important types of EMOs used for learning and teaching is articles. Journal articles published in medical libraries can be harvested and stored as XML

documents using different protocols such as OAI-PMH protocol. One of the widely used libraries by both students and educators are *PubMed Library*. An example of an article harvested from this library is shown in Figure 4.5. The article[8] is harvested using OAI-PMH protocol supported by the PubMed library for publishing its content. A sample from the XML file describing the harvested article is detailed in listing 4.12.



Cardiovasc Diabetol. 2015; 14: 17. PMCID: PMC4328972
Published online 2015 Feb 7. doi: 10.1186/s12933-015-0173-8

Targeting metabolic disturbance in the diabetic heart

Jesús Fuentes-Antrás, Belén Picatoste, Elisa Ramírez, Jesús Egido, José Tuñón, and Óscar Lorenzo

Author information ▶ Article notes ▶ Copyright and License information ▶

This article has been cited by other articles in PMC.

Abstract Go to:

Diabetic cardiomyopathy is defined as ventricular dysfunction initiated by alterations in cardiac energy substrates in the absence of coronary artery disease and hypertension. In addition to the demonstrated burden of cardiovascular events associated with diabetes, diabetic cardiomyopathy partly explains why diabetic patients are subject to a greater risk of heart failure and a worse outcome after myocardial ischemia. The raising prevalence and accumulating costs of cardiovascular disease in diabetic patients underscore the

Figure 4.5: PubMed article example

Listing 4.12: OAI-PMH retrieved record

```
<record><header>
  <identifier>oai:pubmedcentral.nih.gov:4328972</identifier>
  <datestamp>2015-02-15</datestamp><setSpec>cardiab</setSpec>
  <setSpec>pmc-open</setSpec></header>
  <metadata><oai_dc:dc
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/" >
  <dc:title>Targeting metabolic disturbance in the diabetic heart</dc:title>
  <dc:creator>Fuentes-Antrs, Jess</dc:creator>
  <dc:creator>Picatoste, Beln</dc:creator>
  <dc:creator>Ramrez, Elisa</dc:creator>
  <dc:creator>Egido, Jess</dc:creator>
  <dc:creator>Tun, Jos</dc:creator>
  <dc:creator>Lorenzo, scar</dc:creator>
  <dc:subject>Review</dc:subject>
  <dc:description>Diabetic cardiomyopathy is defined as ventricular dysfunction
  initiated by alterations in cardiac energy substrates in the absence of
  coronary artery disease and hypertension...</dc:description>
```

[8] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4328972/

```
17    <dc:publisher>BioMed Central</dc:publisher>
18    <dc:date>2015-02-07</dc:date>
19    <dc:identifier>http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4328972/
20                                  </dc:identifier>
21    <dc:identifier>http://www.ncbi.nlm.nih.gov/pubmed/25856422</dc:identifier>
22    <dc:identifier>http://dx.doi.org/10.1186/s12933-015-0173-8</dc:identifier>
23    <dc:type>Text</dc:type>
24    <dc:rights>This is an Open Access article distributed...</dc:rights>
25  </oai_dc:dc>
26  </metadata>
27  </record>
```

The different types of EMOs demonstrated in this section are mapped to the LEMO metadata schema as explained in the mapping process. All the EMOs are represented as RDF resources in the LEMO metadata schema with predicates for describing its attributes detailed in the original XML files. The mapping process does not cover the full elements exist in the original metadata schema. It maps the proposed LEMO metadata element set with their matching elements in the original metadata elements.

4.7 Summary

Online repositories manage large number of educational resources. It requires having metadata schemas that govern the process of describing its content. Generally, metadata schemas are designed with a purpose in mind. They are developed to satisfy the aims of the organisation building it. This chapter has presented the process of designing, implementing, and testing of the LEMO metadata schema. The metadata schema has been designed as a Dublin Core Application Profile (DCAP) that extended the elements of the DC metadata schema. It has been proposed to comply to the recommendations from the medical education community. Moreover, it addressed the gaps presented in the literature regarding having a simple metadata schema that accommodates the description of different types of educational resources. The LEMO metadata schema has been designed for providing new ele-

ments extending the original DC metadata elements that support the annotation of the textual content of the metadata. It utilises biomedical ontologies for enhancing the description of the medical educational resources to enable its integration. The LEMO metadata schema has been implemented in RDF/XML format to exploit Linked Data for building a one linked dataset of educational medical objects collected from various web data sources.

Before designing the LEMO metadata schema, this chapter has addressed the research objective **O3**: "Conduct a comparative analysis of the existing metadata schemas to identify the common characteristics for describing EMOs". A comparative analysis of existing metadata schemas has identified the common characteristics of used for describing medical educational resources. The rest of this chapter has addressed the research objective **O4**: "Design the proposed LEMO metadata schema by introducing new features to enrich the description of EMOs and enable its integration into one linked dataset. Then, implement the LEMO metadata schema in Linked Data format, and validate that metadata schema by conducting experiments for describing real EMOs of different types that are collected from diverse web data sources". Addressing these two research objectives has answered the research question **R2**: "How can Linked Data practice be used to design and implement a metadata schema that accommodates various types of EMOs and enables its exposing and linking with the aid of external datasets such as biomedical ontologies?". The answer to this research question is the LEMO metadata schema that is implemented in RDF/XML format and tested with educational medical objects of different types such as videos, blogs, and articles. The following section explains the use of this LEMO metadata schema for aggregating and linking distributed Educational Medical Objects (EMOs) into one dataset using the proposed LEMO system.

Chapter 5

Integrating Heterogeneous Web Data Sources

5.1 Introduction

New technologies emerging these days have an immediate impact on all aspects of life in all disciplines, whether commerce, business, health, or entertainment. Recently, Web 2.0 technologies and social networks have been freely used in everyday life. The rapid attitude of embracing new technologies has sparked interests within education. New practices of creating, publishing, and sharing information on the web are becoming the norm these days, replacing the traditional use of the web for consuming information. This change in the web paradigm is predominantly influencing the shape of education as any other discipline. Studies focusing on investigating the potential impact of introducing new technologies in teaching and learning [Bennett et al., 2012][Brown, 2012] [Martin et al., 2011] had motivated this research. The view that Web 2.0 is incorporated into the daily learning and teaching habits of students and educators impose the necessity to consider educational materials of such type in education. Hence, this chapter aims to build a system that can integrate Educational Medical Objects (EMOs) aggregated from distributed web

data sources into one coherent dataset that is easily searchable and accessed. The recommendations resulting from the exploratory study (chapter 3) have stated the frustrations faced by students and educators when searching the web. Moreover, the growing number of heterogeneous metadata schemas used for publishing and organising EMOs advise the necessity of a local repository of metadata records collected from distributed web data sources where dynamic links can be created between EMOs that were not previously linked. Thus, a system that builds on top of the proposed LEMO metadata schema (chapter 4) is presented and referred to as the Linked Educational Medical Objects (LEMO) system. In the LEMO system, several techniques and methods have been developed to utilise the functionalities proposed in the LEMO metadata schema. The output of developing and running the LEMO system is a linked dataset of heterogeneous metadata describing EMOs aggregated from distributed web data sources. The linked dataset is named the LEMO dataset and is built based on the Linked Data practice [Bizer, Heath and Berners-Lee, 2009]. All the EMOs and their related entities are represented in URIs and RDF making the final LEMO dataset machine-readable. The Linked Data practice applied in the LEMO dataset enables linking its content with external datasets such as biomedical ontologies. In the LEMO system, all the necessary methods have been developed for building and enriching the LEMO dataset from distributed web data sources.

5.1.1 Chapter Objectives

This chapter aims to present the LEMO system that is developed to incorporate Linked Data techniques for harvesting, mapping, and linking EMOs from distributed web data sources. The chapter presents the system design that maps to the research objective **O5**: "Establish the LEMO system framework for harvesting, mapping, and interlinking EMOs using Linked Data techniques". The system aims to build a linked dataset named the LEMO dataset by developing several methods for aggregating and integrating EMOs from the web. This chapter elaborates on the methods that

are developed for harvesting and mapping EMOs from web data sources and by that addressing the research objective **O6**. Furthermore, this chapter investigates the biomedical ontologies that can be utilised in the LEMO system in the process developed for enriching the metadata and by that this chapter addresses the research objective **O7**.

The LEMO system focused on bridging the gap between traditional web libraries and Web 2.0 data sources. Hence, experiments conducted in this research included EMOs harvested from the *PubMed Library*, *YouTube*, and *Blogs* as a representative sample of EMOs published on the web.

5.1.2 Chapter Outline

This chapter explains the design and development of the LEMO system and is organised as follows. It begins with a demonstration of the system architecture after explaining the decisions made for proposing this design, followed by an explanation of the main processes and techniques developed in the LEMO system and used for building the LEMO dataset. Finally, experiments and discussions are presented for testing the system on a small dataset. The results of this experiment helped in forming decisions for building the final LEMO dataset aimed for in this research. For example, what web data sources to use for harvesting data and which ontology to use for enriching the content of the LEMO dataset.

5.2 System Design

The LEMO system has been designed to exploit the features introduced in the LEMO metadata schema (chapter 4). The system architecture was designed based on a set of decisions taken to define the LEMO system components, behaviours, and functions. In this section, the LEMO system architecture is detailed after deciding on some architectural decisions that had to be taken to achieve the functional re-

quirements of the LEMO metadata schema, and to satisfy the needs of the medical education domain.

5.2.1 Architectural Decisions

The components constitute the LEMO system arise out of corresponding architectural decisions. The decisions provided clear and decisive guidance on how to proceed with designing the LEMO system. The components of the LEMO system are marked in bold, and they are the results of the following list of decisions. Further elaboration about the implementation of each component is presented later in this chapter.

1. *Rely on existing web data sources*

   The open data movement and the widespread adoption of the Web 2.0 technologies in education resulted in numerous web data sources hosting a large number of quality educational objects. Particularly, relevant datasets provided by the *PubMed Library*, *YouTube* channels, and trusted *Blogs* are the foundation of the LEMO system. These **web data sources** were considered the main input of the LEMO system. This decision was pragmatic since relying on existing educational objects hosted on the web allowed the development of a working system quicker and easier task. In addition to the pragmatic features of this decision, it can be considered a user requirement based decision. The exploratory study conducted with respondents from the domain of medical education had enquired about where they search. The respondents' answers highlighted the use of *PubMed Library* and Web 2.0 data sources when searching for educational objects such as *YouTube* and *Blogs*. Hence, the web data sources decided on considered both popular educational libraries hosting articles and Web 2.0 objects such as videos and blogs. However, this decision results in a heterogeneous dataset because web data sources use different metadata schemas for describing its educational objects. This drawback is

fixed by another decision affecting the development of the LEMO system.

2. *Harvest metadata from the web data sources*

   The set of web data sources decided upon in the previous decision is not limited to a specific number. The LEMO system is designed to have a dynamic dataset that can be extended via **harvesting endpoints** implemented to collect the metadata of content posted on *Blogs* and *YouTube* channels. Additionally, the LEMO system is designed to harvest content from traditional libraries that provides an interface that allow external sources to harvest their content.

3. *Adopt the LEMO metadata schema*

   The EMOs hosted by the selected web database are described in their original metadata schema. Hence, building a linked dataset after harvesting such EMOs requires representing their metadata using a unified metadata schema. Earlier in this research, the necessity to propose the LEMO metadata schema was explained. Experimenting with the LEMO metadata schema proved its capability of accommodating EMOs of various types. Therefore, the decision to adopt the LEMO metadata schema was a systematic decision that was previously decided upon after the intensive research and development presented in Chapter 4. To address the issue of transforming the different metadata schemas into the LEMO metadata schema, a **mapper** was developed to maintain the process of describing all the educational objects using the LEMO metadata schema. The mapping process was implemented as part of the LEMO metadata schema development in Chapter 4.

4. *Semantically enriching the EMOs using ontologies*

   A key objective of the LEMO system was to build a coherent dataset out of the EMOs collected from the web data sources. The LEMO metadata schema has already been developed with features for describing enrichments. The biomedical ontologies are well maintained and increasingly developed as this

research is highly interesting in the medical field. Ontologies are used for annotating various types of medical content, from patient records and laboratory results to medical data and libraries [Hoehndorf et al., 2012]. An existing and well-established annotation endpoint named the **Bioportal annotator API**[1] has been integrated into the LEMO system for annotating parts of the LEMO metadata schema [Noy et al., 2009]. Due to the high number of existing ontologies in the biomedical field, experimenting with different ontologies is needed to decide on one ontology to use in this research. Integrating more than one ontology for annotating text is beyond the scope of the LEMO system at this point.

5. *Store the dataset in RDF triple store*

   Driven by the Linked Data practice for publishing web content [Bizer, Heath and Berners-Lee, 2009], the LEMO system stores the mapped and enriched EMOs in an **RDF Triple Store**. The basic idea of Linked Data practice is publishing structured web data on the web using the RDF data model and linking the distributed data using RDF links [Yu, 2011]. Describing all the EMOs using the LEMO metadata schema that is implemented in RDF standards makes the LEMO dataset machine-readable which enables the automatic interlinking of its content.

6. *Access the dataset using ontologies*

   The final RDF store is machine-readable but not easily understandable for human eyes. Linked Data browsers have been used to navigate the web of data resulted from publishing information on the web using RDF standards [Bizer et al., 2008]. Such browsers can automatically link the documents they navigate. However, these browsers are not widely adopted these days. Therefore, techniques for browsing and querying the linked dataset have been developed

[1] http://bioportal.bioontology.org/annotator

as **access endpoints** based on the ontology used for annotating the data. Simulations of users' behaviour have been implemented in the LEMO system to test the methods developed for accessing the dataset.

The components resulted from the architectural decisions in bold compose the LEMO system architecture.

5.2.2 System Architecture

Based on the architectural decisions discussed in the previous section, the architecture of the Linked Educational Medical Objects (LEMO) system has been designed. The LEMO system architecture is illustrated in Figure 5.1 as a layered structure of processes. All the components of the LEMO system marked in bold in the architectural decisions and their interactions are detailed in this figure.

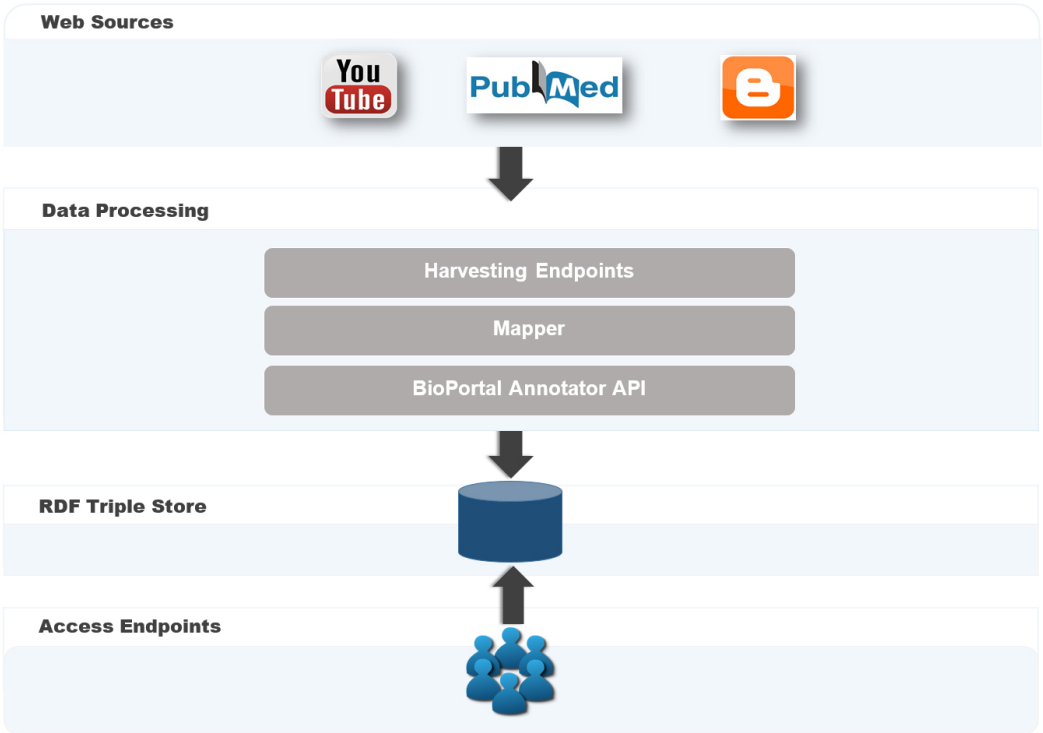


Figure 5.1: LEMO system architecture

The input of the LEMO system is collected from web data sources, and the output of the system is the RDF Triple Store. In between, processing of the data collected includes three tasks: harvesting EMOs, mapping the EMOs into the LEMO AP and annotating the EMOs using external biomedical ontologies. The harvesting endpoint extracts metadata of EMOs from web databases like *PubMed Library*, *YouTube* channels and blogging platforms. The harvesting endpoint retrieves XML files that are stored temporarily in the LEMO system. These XML files store the metadata of EMOs collected which are heterogeneous since they are collected from distributed web data sources. Hence, the mapper handles its transformation into the LEMO metadata schema. It matches the elements of metadata in these XML files into their corresponding elements in the LEMO metadata schema. After representing all the metadata of the EMOs harvested in one unified schema, further processing of the metadata is performed to add semantics to the LEMO metadata elements. The *title* and the *description* of an EMO are enriched with concepts from external ontologies using the BioPortal annotator API. The annotator tool annotates the metadata of an EMO with concepts and represents them as term resources that are linked to the EMO using its LEMO metadata elements. The ontology concepts are organised in a hierarchical structure. The generalization-specialization relations represented in the hierarchy of the ontology can be reflected on the relations between the terms resources describing an EMO. The relation between the term resources is the basis of the links built between the EMOs in the LEMO dataset. Finally, the access endpoints developed exploit these annotations and their hierarchical relations for browsing and querying the dataset.

5.3 System Implementation

This section presents the implementation of the LEMO system architecture and the flow of data in it, starting from distributed EMOs towards building a coherently

linked dataset. The components of the LEMO system, its detailed processes, and their interactions are illustrated in Figure 5.2. The figure demonstrates the process flow in the LEMO system starting from collecting the data from the web and finishing with storing the LEMO dataset in the RDF triple store designed. Additionally, it illustrates the inputs needed for each process to perform its job, and the techniques for transforming the EMOs and integrating them into one linked dataset. A detailed description of the final dataset stored in the RDF triple store is explained in chapter 6. The consequent subsections detail the techniques and algorithms developed for implementing the LEMO system.

5.3.1 Harvesting

The LEMO system presented in figure 5.2 starts with harvesting EMOs from the web data sources using harvesting endpoints. Two endpoints have been developed for this purpose.

- *The RSS feeds reader*: generally, *YouTube* channels and *Blogging* platforms are bundled with RSS feeds that can be read easily using an RSS feeds reader. The developed RSS feeds reader is used to download the RSS feeds as XML files from selected sets of *YouTube* channels and *Blogging* platforms. This endpoint is capable of collecting EMOs published in journal blogs and *YouTube* channels via their RSS feed URLs. Examples of such sources are *Khan academy medicine channel*[2], the blog of *emergency medicine cases*[3], and the blog of *The New England Journal of Medicine (NEJM)*[4]. The RSS feed reader was developed in Java incorporating the ROME API[5] for syndicating RSS feeds. ROME is published as open source under the Apache 2.0 license[6] and is hosted

[2]https://www.youtube.com/user/khanacademymedicine
[3]http://emergencymedicinecases.com/
[4]http://www.nejm.org/
[5]http://rometools.github.io/rome/
[6]http://www.apache.org/licenses/LICENSE-2.0

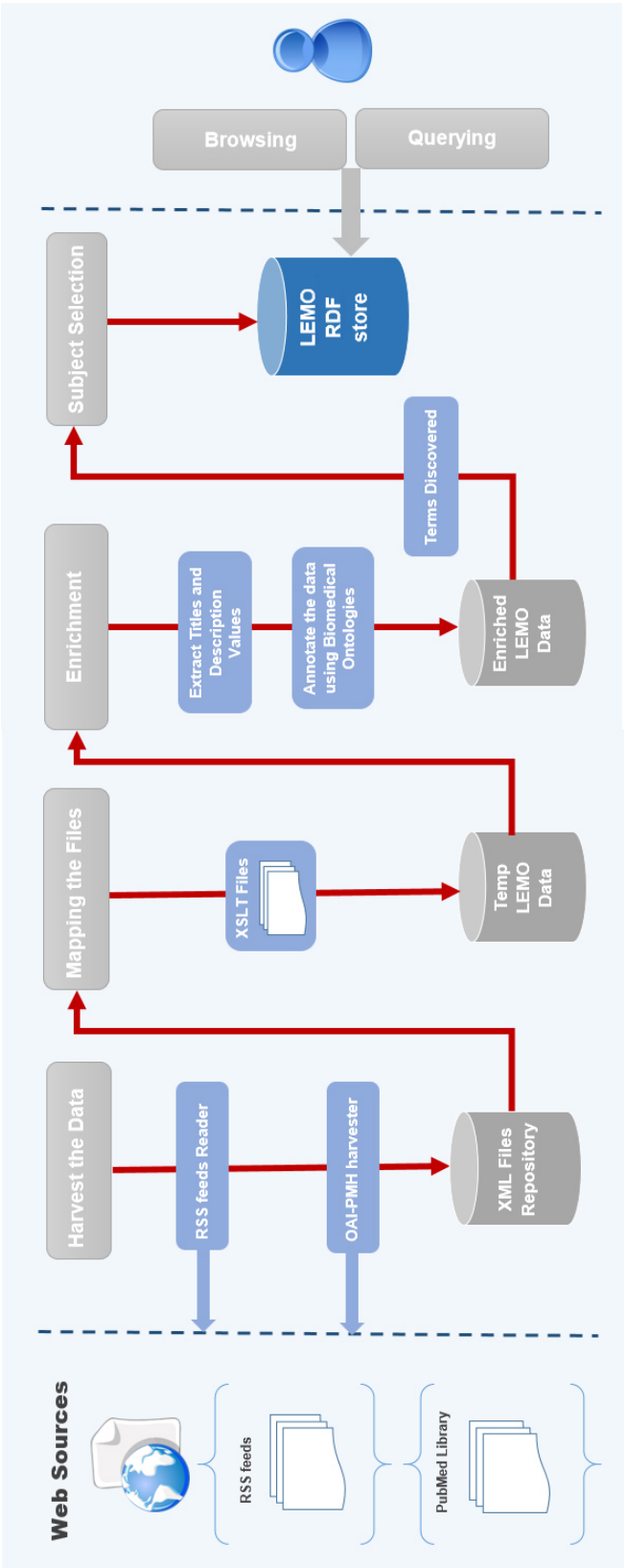


Figure 5.2: Detailed processes and the data flow of the LEMO system

on Github[7] for sharing and access. The ROME API includes parser and generator classes that handle the syndication feeds. The implementation of ROME classes is lightweight and is based on the abstract model of a syndication feed. The input for the developed RSS feeds reader is the RSS feed URLs of blogging platforms and *YouTube* channels. Figure 5.3 illustrate an example of an article published in the New England Journal of Medicine (NEJM). The RSS feeds button highlighted in the orange rectangle at the top right corner of the figure is used to identify the RSS feeds URL used in the harvesting endpoint. As for the *YouTube* channels, the RSS feeds URL is the channel URL and is used by the developed RSS feeds reader to retrieve the videos uploaded by the channel as XML files.

- *OAI-PMH harvester*: this endpoint is an implementation of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) which is a protocol for retrieving metadata from educational libraries [Lagoze and Van de Sompel, 2003]. It is used to collect metadata records from libraries supporting this protocol such as the *PubMed Library*. Few of the educational libraries provide open content for their users, and even fewer libraries provide the ability for others to harvest and store their data such as the *PubMed Library*. The PubMed Library is set up to use this protocol that provides access to its metadata. The OAI-PMH harvester is used to collect articles and store their metadata as XML files in addition to the RSS feeds collected from the first endpoint.

After harvesting data using these two endpoints, a collection of heterogeneous XML files is stored in an XML file repository for further processing. These files are the entry for the second process that is responsible for mapping the XML formats into the LEMO metadata schema.

[7] https://github.com/rometools/rome

Figure 5.3: RSS feeds button for NEJM blog article

5.3.2 Mapping

The XML files repository stores all the metadata collected via the harvesting endpoints. The following process in the LEMO system maps these XML files into one unified schema that is the LEMO AP schema. As part of developing the LEMO metadata schema (Chapter 4), the process of mapping any XML file into the RDF/XML format of the LEMO metadata schema has been explained. The mapping process is performed based on XSLT templates designed for transforming the different XML formats stored in the XML files repository into the RDF/XML format of the LEMO metadata schema. This mapping process has been implemented at an earlier stage of the research (Chapter 4, section 4.5.2) to ensure that the LEMO metadata can be used to describe various types of EMOs as specified in its functional requirements.

After the mapping process, all the XML files are transformed into the desired RDF/XML format of the LEMO metadata schema and stored in a temporary

RDF triple store. Usually, the metadata provided with *YouTube* videos or Blogs are less descriptive than the metadata of EMOs harvested from educational libraries such as *the PubMed Library*. The reason is that content published on Web 2.0 data sources is not constrained by rules and regulation of filling its metadata with proper descriptions. The metadata of Web 2.0 data is user-generated while the metadata describing EMOs in online educational libraries is maintained by the publishing organisation. Hence, the metadata of EMOs collected from online educational libraries is usually of a higher quality than Web 2.0 EMOs. The metadata quality and completeness are considered one limitation of this research, especially for videos and blogs metadata. Therefore, all the EMOs metadata collected in the temporary RDF store are the entries for the next process of the LEMO system that handles enriching the metadata elements with further semantics.

5.3.3 Enriching

In this process, the LEMO system presents methods for enriching the metadata of EMOs by adding semantics to their elements. First, EMOs titles and descriptions are annotated with concepts from the ontologies. These concepts are represented as Term resources related to the EMOs in the LEMO metadata schema. Further processing of the Term resources can filter the terms to a smaller and more descriptive set of Term resources that represent the subject attribute of EMOs. Thus, the LEMO dataset can be interlinked based on the relations between the Term resources and can be categorised based on their subject attribute. In order to understand the enrichment process, it is necessary to review the LEMO metadata element set illustrated in figure 4.2 explained in the metadata implementation phase (Chapter 4, Section 4.5).

Before the enrichment process, all the EMOs are described in LEMO metadata schema using the DC-based elements resulted from the mapping process that matches the values in the original metadata elements of each EMO. In other words,

each EMO at this stage is described with RDF statements using the predicates from the (dc:) prefix, while the RDF statements that uses the prefix (lemo:) are not yet created. The enrichment process is responsible for adding new resources and creating the necessary RDF statements that handle enriching the EMOs. Hence, the process of enriching an EMO is split into two sub-processes: terms discovery and subject selection.

Terms Discovery

The annotation tool developed in the LEMO system uses well-established biomedical ontologies for adding further refinements to the EMOs metadata. Ontologies are formal representations of knowledge with the definition of concepts and their relations [Rubin et al., 2008]. Ontologies have been used in libraries for indexing entries and ease the search process, such as the use of MeSH ontology for indexing PubMed library entries [Jonquet et al., 2011]. The LEMO system implemented the annotation tool as part of the enrichment process by incorporating the BioBortal Annotator API[8] which is provided by the BioPortal repository as a web service [Noy et al., 2009]. The API takes as arguments the text to be annotated and the ontology name to be used in the annotation process, and returns the annotated text in JSON format. The concepts of the ontologies in the BioPortal repository are represented as class instances and the relations between these concepts are organised in a class hierarchy that stores the taxonomic (is-a) hierarchy where concepts may have multiple parents.

In the LEMO system, the annotation process is explained in algorithm 1. For each EMO described in LEMO metadata schema, the text to be annotated is extracted from the value attribute <rdf:value> of both the Title and the Description resources that are related to that EMO via <dc:title> and <dc:description> predicates respectively.

[8]http://bioportal.bioontology.org/annotator

Algorithm 1 Terms Discovery of EMOs

```
1: Input : The temp LEMO store
2: Output : A collection of Term Resources and its related Class resources
3:
4: tempStore ← get the Title and the Description resources from temp LEMO
   store
5: procedure AnnotateText(tempStore, OntologyName)
6:     for all r in tempStore do
7:         v ← get literal object of <rdf:value> for r
8:         result ← getAnnotations(v,OntologyName)
9:         terms ← parse(results)
10:        termResources ← createRDFtermResources(terms)
11:        add termResources to tempStore
12:        add the necessary predicate to link r to its matching termResources.
13:        UpdateLEMOGraph(termResources, OntologyName)
14:    end for
15: end procedure
16:
17: procedure UpdateLEMOGraph(termResources,OntologyName)
18:    for all t in termResources do
19:        classID ← getClassID(t)
20:        c ← createClassResource(classID)
21:        getPathToRoot(classID)
22:        update attributes of c
23:        if c not in tempStore then
24:            add c to tempStore
25:        end if
26:    end for
27: end procedure
```

As explained in algorithm 1, The LEMO system annotates the literal values of the EMOs described in the Title and the Description resources related to each EMO via the BioPortal API. Then, the annotation results are read and parsed in this process to create the Term resources representing these annotations. Each annotation retrieved tags the text with concepts from the ontology used in the API request. The Term resource (Chapter 4, listing 4.5) describes the indices of the text that has been annotated, the concept it has been annotated with using the class ID of that concept, the class name, the synonym, and the definition if it exists. Then, the necessary RDF statements are added to the Title resource (`<lemo:lemoTitleAnnotations>`), and the

Description resources (<lemo:lemoDescAnnotations>) pointing at the terms discovered in its value text predicate (<rdf:value>). Also, Class resources are created, and their taxonomic relations are retrieved from the BioPortal repository via API requests that return the path from the class requested to its root. The Class resources relations are represented by adding RDF statements that represent the adjacency predicates <lemo:adjacentTo>. The annotation tool results in a large number of annotations represented as Term resources and added to the LEMO RDF store. The class resources and their relations are updated consistently with each annotation added to the LEMO RDF store forming a smaller version of the ontology taxonomic hierarchy named the LEMO graph. The collection of terms related to each EMO can be used for categorising the EMOs as explained in the second process of the enrichment.

Subject Selection

The subject of an EMO is modelled in the LEMO metadata schema using the DC predicate <dc:subject>. Each EMO can be categorised using this LEMO metadata element by relating it to the most relevant terms discovered in that EMO resource after filtering them. The next step of the enrichment process is to filter the terms discovered in each EMO based on the subject selection algorithm detailed in algorithm 2, and relate the most relevant Term resources as objects for the predicate <dc:subject> that stores the subject of that EMO.

Each term annotated and related to an EMO is weighted according to its number of appearances in the Title or the Description Resource. According to studies in the information retrieval field, terms found in the title of any piece of information are useful indicators of the category to which this piece of information belong [Van Hage et al., 2004]. Hence, terms annotated in the Title resources of the EMOs are weighted with higher values than terms annotated in the Description resources. Also, the weights are updated if the same term was annotated more than

Algorithm 2 Subject Selection of EMOs

1: **Input** : The set of Term Resources and its related Class Resources
2: **for all** Resources r as EMO Resource **do**
3: EMOTerms $t \leftarrow$ getTitleAnnotation(r)
4: $t \leftarrow$ getDescAnnotation(r)
5: **for** $t_i \in t$ **do**
6: weight $\leftarrow$ weightBasedonOccurence(t_i)
7: weight $\leftarrow$ weightBasedonHierarchy(t_i)
8: assign weight to the `<rdf:value>` of t_i resources
9: **end for**
10: normalizeWeights(t)
11:
12: subjects $\leftarrow$ sortAndSplit(t)
13: create the necessary RDF statements related to r for describing the subjects using `<dc:subject>` predicate.
14: **end for**

once in the EMO title or description text accordingly. The weights are stored in the `<rdf:value>` attribute of the Term resource. Another factor affecting the term's weighting is its hierarchical position in the ontology. Since the terms are related to Class resources that store the classes hierarchy using its adjacency attributes, the position of the class in the LEMO graph hierarchy affects the weight of the term related to that class. Therefore, the weights of the terms are updated in the subject selection process. Each set of terms annotated for each EMO can be represented as a graph structure where relations between the terms are deduced from the classes adjacency attributes. The weight of each term related to each EMO is then updated to the accumulated weights of its descendent terms based on the adjacency relations of its related classes. Updating the weights of the terms to consider their hierarchical positions in the ontology is beneficial for categorising the EMOs into subjects. Doing so, terms with classes that are leaf nodes in the ontology will have lower weights compared with terms related to classes in higher levels of the ontology. The final step in selecting subjects for EMOs is to normalize the weights of the terms related to EMOs Title and Description resources and specify a threshold for splitting the terms into two sets. The set of Term resources having the highest weights are linked

to the subject predicate <dc:subject> of the EMO it describes by creating RDF statements that update the EMO resource.

The goal of annotating EMOs using ontologies is to build relations between these EMOs. Based on the terms created to represent annotations of the Title and the Description resources, linkages can be generated between EMOs based on the relations between its Term resources. Having the same ontology class annotated by two terms discovered in two different EMOs indicate a link between these two EMOs. However, such links are not valuable for a large number of annotations discovered in large datasets. Therefore, to generate stronger linkages between EMOs in the LEMO dataset, a link is considered to be valid between two EMOs if they have at least one ontology class in common based on the terms related to that EMO via the subject <dc:subject> predicate. Links between EMOs based on its annotation terms discovered either in their Title or their Description resources are still considered as links between EMOs but, in this work, the only valid links are considered between EMOs based on their subject elements.

The following section details the experiments conducted to test the LEMO system with real collected from web database and test the annotation and linking process via two well-known biomedical ontologies: MeSH and SNOMED CT.

5.4 Experiments and Discussions

In this section, a preliminary experiment has been conducted with a sample dataset for testing the functionality of the LEMO system and its integration using different biomedical ontologies. The decision regarding which web data sources to use in the the final LEMO RDF store is taken based on the results of this preliminary experiment. Additionally, the decision of which ontology to be applied for enriching the final LEMO RDF store is taken based on the comparison between the MeSH SNOMED CT annotations detailed in this section.

5.4.1 Dataset Harvested

The harvesting endpoints developed in the LEMO system were used to collect the dataset for this experiment. The dataset consists of videos published in medical *YouTube* channels and blogs managed by medical journals, both harvested using the RSS feeds reader endpoint. Additionally, the dataset consists of articles harvested from the *PubMed Library* via the OAI-PMH harvesting endpoint. The details about the dataset involved in this experiment are detailed in Table 5.1.

Table 5.1: The components of the first LEMO dataset experiment

EMO type	number of EMOs
Article	1000
Video	1259
Blog	461
Total	2720

5.4.2 Biomedical Ontologies

The LEMO system application domain is medical education. Hence, the BioPortal[9] open repository for biomedical ontologies was used to explore the ontologies to use in enriching the LEMO dataset. In this experiment, the LEMO system experimented with these two ontologies: the Systematized Nomenclature of Medicine - Clinical Terms **SNOMED CT** and the Medical Subject Headings **MeSH**.

SNOMED CT

Specialized organisations in both USA and the UK has been developing and maintaining the SNOMED CT ontology. It offers a standardized healthcare terminology that is comprehensive and scientifically validated. The ontology provides a clear description of concepts and the relationships built between them [Elevitch, 2005]. The SNOMED CT was released in 2002 and since then new versions of it have

[9] http://bioportal.bioontology.org/

been released semi-annually. This ontology is popular in the domain of medicine applications. It has been designated as the preferred clinical terminology to use in 19 countries [Lee et al., 2014] and its application in medical information systems is expected to increase. Therefore, the decision to use this ontology for testing the LEMO system functionality is based on its popularity and wide usage of this ontology in research.

MeSH

The MeSH ontology has been developed by the National Library of Medicine (NLM) in the US. The concepts defined in this ontology were used as the controlled vocabulary applied for indexing the content of the *PubMed Library*. It is represented as a set of concepts organised in a hierarchical structure or taxonomic hierarchy that describes the relations between the concepts from general to more specific concepts [Lipscomb, 2000]. In this experiment, EMOs have been harvested from the *PubMed Library* and since the MeSH ontology is used for indexing content, the MeSH ontology is incorporated in the LEMO system for testing its functionality.

5.4.3 Annotations

The enrichment process in the LEMO system consists of two steps: firstly, the term discovery process and secondly, the subject selection process. The results of these process are a collection of Term and Class resources added to the LEMO RDF store. Linkages between the EMOs can be deduced from the relations between the Class resources relating to these Term resources discovered. In this section, the usage of the two biomedical ontologies, MeSH and SNOMED CT, is tested. The results of the term discovery process are compared to decide on which ontology to use for further building the final LEMO dataset. The number of Term resources created and annotated in either the Title or the Description resources using the MeSH, and SNOMED CT ontologies are detailed in Table 5.2. With the proper ontology

Table 5.2: Number of terms annotated for the set of EMOs using different ontologies

		MeSH annotations			SNOMED CT annotations		
Type of EMOs	Number of EMOs	Title	Description	Total	Title	Description	Total
Article	1000	3887	12192	16079	6166	29859	36025
Video	1259	3027	4304	7331	3677	5710	9387
Blog	461	754	4720	5474	1572	9756	11328
Total	2720	7668	21216	28884	11415	45325	56740

alignment methods, the two ontologies can complement each other when annotating text. Currently, this research does not support ontology alignment, but further details about the applicability of such techniques are presented as future work of this thesis.

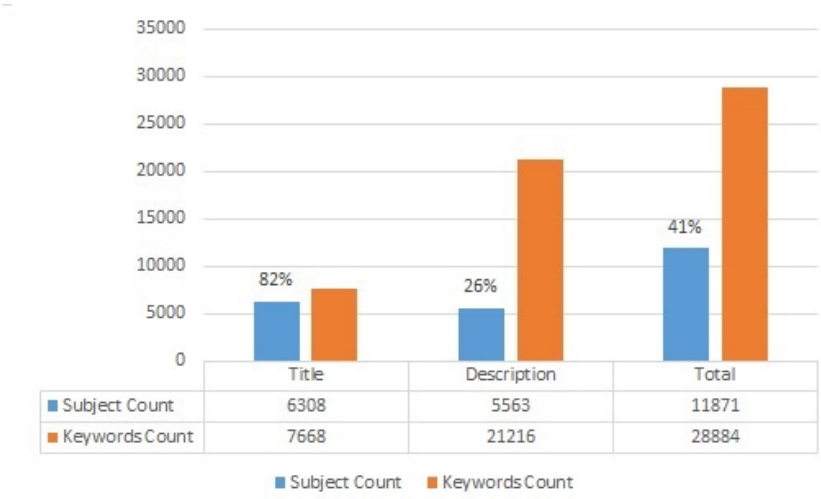
It can be noticed that the number of the Term resources annotating the EMOs using the SNOMED CT ontology is greater than the number of Term resources annotating the EMOs using the MeSH ontology. The difference is not significant for EMOs of type videos and blogs when compared to the EMOs of type articles. This difference is due to the short textual description provided in the metadata of blogs and videos compared to the longer text provided for articles in the online libraries. The number of the terms discovered affects the next process of subject selection. The selection of Term resources to represent the subject attribute of an EMO is tested against the two discovered sets of terms in both ontologies.

5.4.4 Subject Selection

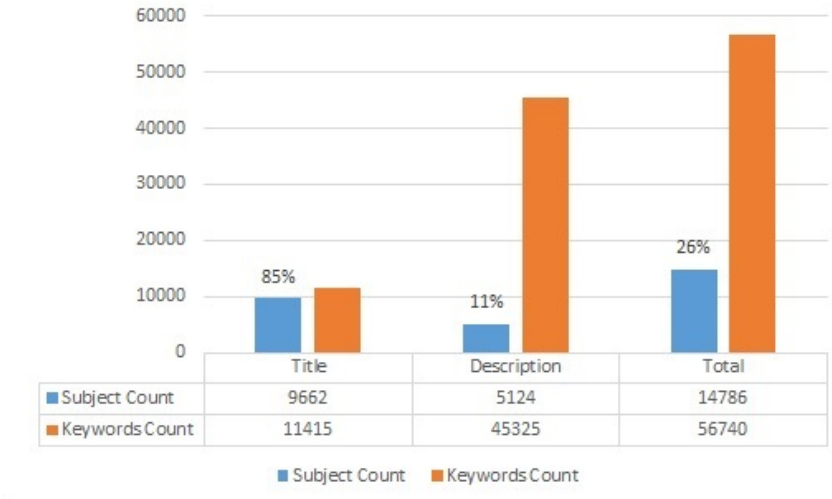
After the term discovery process that annotates the titles and descriptions of EMOs with classes of ontologies, processing the set of terms related to each EMO results in a weighted set of terms that represents them in descending order from the most important terms to the less important ones. The values of the weights are dependent on two important factors: the number of occurrences of the terms in the Title or the Description resources and the position of the class ontology annotated by the

Term resource in the ontology. The detailed process of subject selection has been explained in algorithm 2. The Term resources with the highest weights are selected as terms that represent the subject of the EMO. A threshold value is calculated based on the weights and number of terms discovered for each EMO. Hence, the count of Term resources related to the subject attributes `<dc:subjet>` of each EMO is variable and depends on the terms discovery process.

In this experiment, the total number of Term resources selected as subject



(a) Subjects based on MeSH



(b) Subjects based on SNOMED

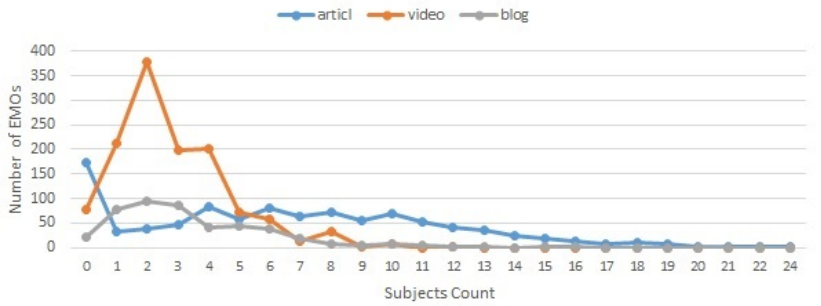
Figure 5.4: Comparison of subject selection process using two ontologies

attributes in the LEMO dataset is compared to the total number of Term resources related to the Title or the Description resources. The results are detailed in figure 5.4a and figure 5.4b for the MeSH and SNOMED CT subject selection process respectively.

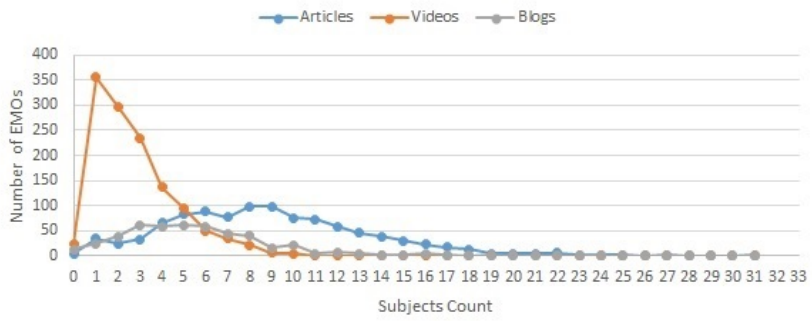
The percentage of Terms resources selected as subjects for EMOs using the SNOMED CT set of terms is less than the percentage of the MeSH terms annotated. In both experiments, the subjects selected are mainly chosen from the Term resources related to the Title resources of EMOs. Since the title of any object on the web gives a better indicator about the subject of that object, the results indicate that the subject selection algorithm has succeeded in this matter. The results illustrate that, using the SNOMED CT ontology, only 11% of the Term resources annotated in the Description resources were chosen as a subject attribute, compared to a higher percentage of 26% Term resources annotated in the Description resources discovered using MeSH.

The LEMO dataset consists of different types of EMOs. Usually, video and blog EMOs have less descriptive metadata content. Hence, the number of Term resources discovered and related to the EMOs is affected by the types of EMOs, and that reflects on the number of subjects selected. Using MeSH and SNOMED CT annotations, the results of selecting subjects for EMOs of type videos and blogs is similar, and no improvement is shown when using different ontologies. Figure 5.5a and Figure 5.5b illustrate the relation between the number of subject attributes added to EMOs and their types after the subject selection process applied using MeSH and SNOMED CT annotations.

In both experiments, video and blog EMOs have a small number of subject attributes added. This is due to the low numbers of terms annotated in this type of EMOs. The comparison results of the subject selection process between the MeSH and the SNOMED CT showed that using the SNOMED CT annotations, very few EMOs did not have any Term resource selected as a subject. While using the MeSH



(a) EMOs annotated in MeSH and their subjects



(b) EMOs annotated in SNOMED and their subjects

Figure 5.5: Relation between subjects count and EMOs types

annotations, more than 150 EMOs from articles, videos or blog types have zero Term resources selected for representing its subject attribute and that affects linking the content of the LEMO dataset.

5.4.5 Linkages

After the enrichment process, all the EMOs can be represented by a set of terms annotated with concepts from biomedical ontologies. These concepts are part of a larger hierarchical taxonomy where all the concepts are organised in one ontology. In the term discovery algorithm (algorithm 1), the hierarchical relations of each term discovered, while annotating the EMOs, are retrieved and stored in the Class resource description. The relations between the ontology classes annotated in the Term resources form a sub-graph of the original ontology. Hence, relations between

Table 5.3: Links in LEMO dataset based of MeSH and SNOMED CT ontologies

	Number of Links based on		
Based on Ontology	Title Annotations	Description Annotations	Subjects
MeSH	352029	867636	248704
SNOMED CT	464876	1443667	418782

the Class resources indicate a relation between the EMO resources that are connected to these Class Resources via the LEMO metadata elements describing the Title and Description resources. Usually, if two pieces of information contain similar parts of the text, then the two pieces of information are somehow similar. Therefore, in the LEMO system, EMOs are related to each other if they are annotated by the same Class resources. In other words, a link is automatically generated between two EMOs if they point to the same Class resource based on the Terms resources annotating its Title or Description resources. A link between two EMOs is considered as two directed links. Therefore, if there is a link from node a to node b then the number of links between these nodes is two links. After this clear definition of what is a link in the LEMO dataset, the detailed statistics of links generated in this experiment is presented in Table 5.3. It compares the number of links generated in the LEMO dataset based on the experiments of enriching its content using two different ontologies.

The table illustrates the number of links generated in the LEMO dataset based on different attributes of EMOs: its title annotations represented in the Title resources `<lemo:lemoTitleAnnotations>`, the description annotations represented in the Description resources `<lemo:lemoDescAnnotation>`, and the subject attribute `<dc:subject>`. The results indicate that the number of links generated based on the SNOMED CT ontology enrichment is greater than the links generated based on the MeSH ontology enrichment. The results are almost doubled in the links count. This is due to a large number of annotation discovered using SNOMED CT ontology.

5.5 Summary

This chapter has presented the LEMO system that was proposed to harvest, enrich, and store Educational Medical Objects (EMOs) from distributed web data sources into one linked dataset. The system has been developed to utilise the features of the LEMO metadata schema proposed in the previous chapter. It consists of processes that are responsible of collecting EMOs from the the web, map its metadata into the LEMO metadata schema, and enrich its description with concepts from biomedical ontologies. This chapter has examined the use of two different biomedical ontologies, MeSH and SNOMED CT, for annotating a dataset of 2720 EMOs collected from *YouTube*, blogs, and the *PubMed library*.

From the analysis of the results of these experiments, the system has proved its ability enrich the metadata of different EMOs described in the LEMO metadata schema. The experiments that has been conducted to test the use of different biomedical ontologies has shown that SNOMED CT ontology produced more annotations to the EMOs in the dataset. Such results has helped to decide on using SNOMED CT ontology in the LEMO system for enriching further EMOs harvested. The links generated based on the SNOMED CT ontology almost doubled those generated by MeSH ontology annotations. The goal of the LEMO system was to build a linked dataset of distributed EMOs collected from the web, and the SNOMED CT presented better results for achieving this goal.

In conclusion, this chapter has addressed the research objective **O5**: "Establish the LEMO system framework for harvesting, mapping, and interlinking EMOs using Linked Data techniques". Furthermore, this chapter has presented the work implemented for addressing the research objectives elicited for building the LEMO system framework, that are research objectives **O6** and **O7**. The work resulted from achieving all these research objective has answered the research question **R3**: "What are the techniques used for harvesting, mapping, and organising EMOs from

distributed web data sources into one linked dataset?". The main outcome of this chapter is developing the LEMO system and deciding on using the SNOMED CT ontology for enriching the EMOs in order to have one linked dataset named the LEMO dataset. The detailed description of the final LEMO dataset resulted from this LEMO system is explained in the following chapter.

Chapter 6

The RDF Triple Store

6.1 Introduction

This chapter presents the final results after running the LEMO system for aggregating and integrating Educational Medical Objects (EMOs) in one linked dataset named the LEMO dataset. The detailed process of building the LEMO system has been explained in the previous chapter (Chapter 5). The LEMO system exploits the features presented in the LEMO metadata schema (Chapter 4) that is proposed to solve the problem identified earlier in this research (Chapter 1, section 1.1). Based on the experiments that have been conducted in the development phase of this research, the work has been extended to include a wider range of EMOs aggregated from distributed web data sources and annotated with concepts from the SNOMED CT ontology. The results of testing the system and comparing the results of incorporating different ontologies and data sources have guided the decisions to extend the LEMO dataset and annotated its content with SNOMED CT ontology. The LEMO system developed aimed at building a linked dataset of EMOs harvested from distributed web data sources named the LEMO dataset. The web data sources host EMOs of various types such as articles, videos, and blogs. The final LEMO dataset is stored in an RDF store since the metadata of the EMOs are described using

the LEMO metadata schema that is developed in RDF/XML format. This chapter describes the content of the RDF store storing the LEMO dataset and details the results of enriching and thus annotating its components. The RDF store is accessed via ontology-based techniques developed in the following chapters for browsing and querying its content. Thus, the description of the RDF store content is essential for understanding the evaluation of the LEMO dataset by testing information retrieval from the RDF store via browsing (Chapter 7) or query searching (Chapter 8).

First, this chapter summarizes the terminologies used to describe the RDF store, and it introduces a formal description of the RDF store components in mathematical notations that will be used to describe the evaluation techniques developed in the following chapters. The following list summarizes the terminologies and their synonyms that are used to describe the content of the RDF store, often called a triple store, in order to enhance the readability of this chapter.

- *An RDF statement:* is a statement made up by an RDF triple in the form of a subject-predicate-object expression. A triple can be represented as an RDF graph that connects the subject and object with a predicate that defines how they are related.

- *Subject:* is a resource in an RDF store that is identified by a URI and described using predicates. It is considered the start node in an RDF graph.

- *Predicate:* is an attribute that defines the relation between the subject and the object of an RDF statement and it represents an edge in the RDF graph. It is identified by a URI that is translated from a prefix that identifies the namespace of the vocabulary to describe the RDF statements.

- *Object:* is the value of the predicate describing a subject resource. It can be a resource or a literal value. Usually, it is considered the end node in the RDF graph representing an RDF statement, but in the case where the object is an RDF resource it can be a start of another RDF graphs. In other words, the

object resource can be described with RDF statements making it a subject resource of these statements.

- *Literal value:* is a value that represents the predicate value describing a subject. It can be a string, URL, or even a numeric value.

- *LEMO metadata schema:* is the metadata schema proposed in Chapter 4. It can be referred to as LEMO AP since it was implemented as a DCAP.

Experiments that have been discussed while developing the LEMO system (Chapter 5, section 5.4) compared the use of two different ontologies: MeSH and SNOMED CT. The number of annotations and linkages created based on testing these two ontologies for enriching the LEMO dataset showed that SNOMED CT was able to annotate more terms and thus generate more linkages in the LEMO dataset. This research does not support the use of multiple ontologies for annotating its content. Therefore, deciding on using the SNOMED CT ontology is based on its ability to generate richer annotations when incorporated in the LEMO system. Moreover, enriching the LEMO dataset with annotations is the basis for building linkages between its content. The more annotations discovered in the EMOs metadata, the more linkages can be built between these EMOs. The EMOs are stored in an RDF store that organises their related resources to generate linkages that are responsible for integrating them. Understanding the organisation of the LEMO dataset in this RDF store is necessary to understand the techniques developed for accessing it. Hence, this chapter presents a scenario of one EMO harvested from the web, and it illustrates how it is stored in the RDF store.

6.1.1 Chapter Objectives

This chapter addresses the research objective **O8**: "Describe the RDF store that is managed by the LEMO system for organising the EMOs metadata represented in the LEMO metadata schema". The main objective of this chapter is to explain how

the LEMO dataset is organised in the RDF store by providing a detailed description of the RDF resources composing the metadata of the EMO in the scenario section. Thus, providing a practical example of the RDF resources that were explained while developing the LEMO metadata schema (Chapter 4, section 4.5). Furthermore, this chapter provides a formal representation of the LEMO dataset components that will be used in the rest of this chapter and the following chapters for referencing the RDF resources in the LEMO dataset.

6.1.2 Scenario Example

The EMO, illustrated in Figure 6.1, is an example of an article retrieved from the *PubMed Library*[1]. This example EMO is described with the LEMO metadata

[1]http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2820004/

Figure 6.1: An example of an EMO retrieved from the *PubMed Library*

schema and enriched using the SNOMED CT concepts. The EMO is described by a collection of related RDF resources that represent its metadata as RDF statements. Annotations discovered after enriching the EMOs in the LEMO system are represented as RDF resources and added to the RDF store as explained in the previous chapter. Figure 6.2 details a sample of the RDF resources describing this EMO as

Figure 6.2: Sample of the EMO metadata schema

stored in the RDF store. The figure illustrates three types of related RDF resources describing the metadata of the EMO in the figure. The EMO resource with the URI (`rdf:about="oai:pubmedcentral.nih.gov:2820004"`) and a Title resource with the URI (`rdf:about="oai:pubmedcentral.nih.gov:2820004:title"`) that represent the object value in the RDF statement described using the predicate `<dc:title>`, and the third RDF resource represents a Term resource that is related to the two resources has the URI (`rdf:about="oai:pubmedcentral.nih.gov:2820004:title:term6"`). This Term resource is considered the object for two RDF statements: one describing the EMO resource via the predicate `<dc:subject>` and the other describing the Title resource via the predicate `<lemo:lemoTitleAnnotatioin>` that defines the annotation discovered in the EMO resources. All the relations between the RDF resources in this figure are illustrated using the red arrows for enhancing the readability of this LEMO metadata sample.

The sample of the RDF/XML snippet in Figure 6.2 can be read easily from an RDF graph of nodes and edges. The nodes are either RDF resources represented as the rectangles, or literal values, and the edges connecting these nodes are the predicates. Figure 6.3 represent an RDF graph describing some of the RDF statements in figure 6.2.

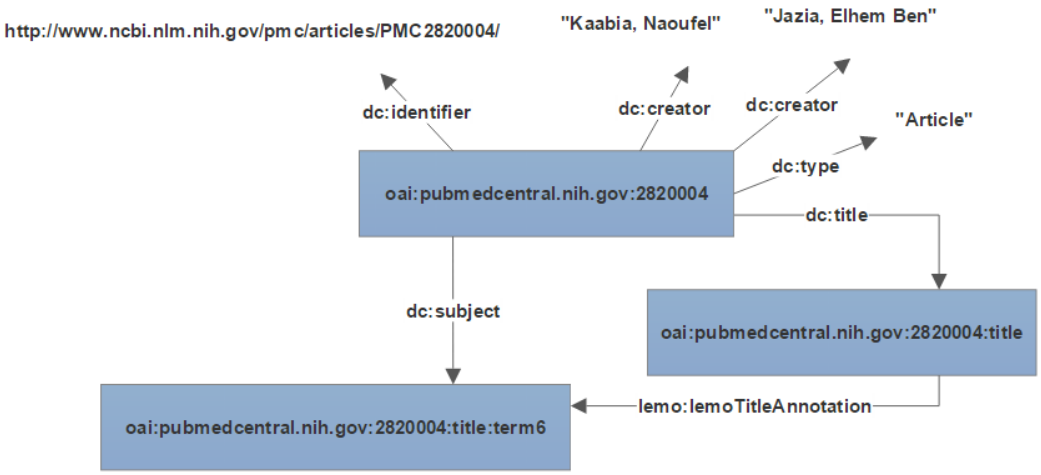


Figure 6.3: An RDF graph representing relations between the RDF resources

6.1.3 Chapter Outline

This chapter begins with an explanation of the notations used to represent the LEMO dataset components. Next, the RDF store content is described giving the detailed numbers of EMOs comprising the LEMO dataset, the RDF resources added for enriching the EMOs, and the number of links generated in the RDF store.

6.2 The RDF Store Implementation

The LEMO system is responsible for aggregating EMOs from diverse web data sources, mapping their metadata into the LEMO metadata schema, enriching its description with SNOMED CT concepts, and store the EMOs metadata records and their enrichment in an RDF store. These processes of harvesting, enriching and storing EMOs metadata are performed offline. Once the dataset is ready and stored in an RDF store, it can be published online for users to access. At this stage of the research, the techniques proposed to build the LEMO dataset and access have to be evaluated to prove its efficiency and provide reasonable answers for the questions initiating this research. Once the results are validated and its effectiveness is proved, the LEMO dataset can be published for public access via a web-based application that is the future research intended for this work. Prototypes of the web-based application were developed for accessing the LEMO dataset as will be illustrated in the following chapters. In this section, the technique used for implementing the RDF store as part of the LEMO system is explained after illustrating the organisation of all the RDF resources in the RDF store.

The RDF store components are detailed in Figure 6.4. The RDF store consists of RDF resources described using predicates to represent the EMO metadata. The RDF resources and their relations are illustrated in a layered design that is intended to define the collections of resources need to describe an EMO. The main resource is the EMO resource at the top layer and the smallest resource shown at

the bottom layer is the Class resource, and the resources in between are added to enhance the description of the EMO resource. The bottom layer is a collection of Class resources that represent the SNOMED CT concepts used to annotate the EMOs and its hierarchical relations that are retrieved from the ontology.

The **first layer** consists of EMO resources which are described using the LEMO metadata schema elements as shown in the RDF/XML snippet. The object values of the two predicates `<dc:title>` and `<dc:description>` are Title and Description resources composing the **second layer**. These resources are described using predicates that identify the annotations discovered in their textual values as shown in the XML snippet. All the annotations discovered in all the Title and Description resources are represented as Term resources comprising the **third layer**. The Term resources are the objects in the RDF statements describing the Title and Description resources. Each Term resource represents a concept of the SNOMED CT ontology used to enrich the LEMO dataset. The XML snippet of the Term resources represents the predicates used to describe the part of text annotated and the concepts it is annotated to as retrieved from the SNOMED CT ontology. Each

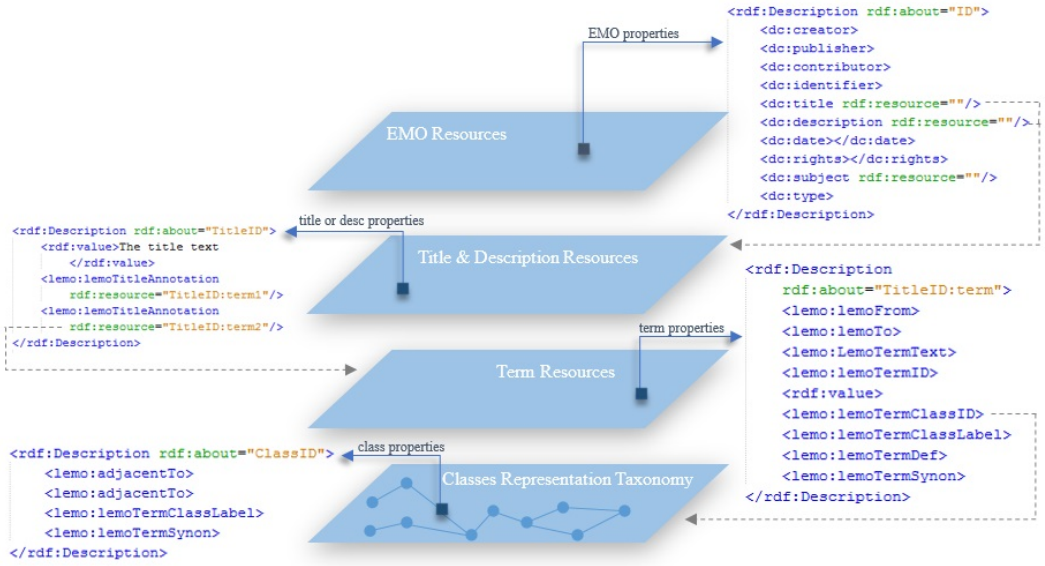


Figure 6.4: Layered design of LEMO dataset components

concept is identified by an ontology class URL that is stored in the Term resource description. The collection of Class resources representing these concepts form the **fourth layer** in the figure. In this layer, the hierarchical relations of the ontology classes are retrieved as structured in the SNOMED CT ontology and these relations are described in the Class resources predicates. Whenever a new class is used to annotate a Term resource in the LEMO dataset, a new Class resource is created, and the relations between the Class resources are updated.

Existing platforms for RDF data management such as *Virtuoso*[2] exist for managing RDF data and hosting it on the web. However, the LEMO system implemented the RDF store using APIs provided by the Apache Jena framework[3]. The processes responsible for mapping and enriching the EMOs, in the LEMO system, after harvesting them create RDF statements that describe these EMOs. Moreover, the LEMO system maintains the hierarchical relations between the Class resources by updating their descriptions. Therefore, constant access to the RDF store is required to add the necessary RDF statements and update the descriptions of others. So, the RDF store is created by the LEMO system from scratch to accomplish the goal of integrating distributed web data sources into one linked dataset and online platforms such as Virtuoso were not efficient to use for building the LEMO dataset.

The Apache Jena framework is an open source Java framework for building Linked Data applications. Using the RDF API provided by Jena, the LEMO system created RDF resources needed for describing the EMOs using the LEMO metadata schema. Creating the RDF resources and writing the RDF statements for describing these resources results in a collection of RDF resources that are stored in the LEMO RDF store. The storage and organisation of these RDF statements are managed using the TDB API provided by Jena. This component of the Jena framework is used for RDF storage and query. A TDB can be used as a high-performance RDF store on a single machine, and since the LEMO store is not published yet for user

[2]http://virtuoso.openlinksw.com/
[3]http://jena.apache.org/

access, this API is beneficial for preparing the LEMO dataset. These two APIs (RDF API, and TDB API) provided for managing an RDF triple store are incorporated in the LEMO system developed for building the LEMO dataset. Further techniques have been developed for accessing and evaluating the content of the LEMO RDF store using the TDB API from the Jena framework.

The ensuing subsections explain a sample of the RDF statements stored in the RDF store to describe the EMO illustrated in the scenario section (figure 6.1). This sample represents a practical example of the RDF resources needed to describe an EMO as explained while developing the LEMO metadata schema.

6.2.1 The EMO Resource

An RDF resource describes the EMO metadata as RDF statements with predicates that represent the LEMO metadata schema as shown in listing 6.1. The RDF statements are triples of subject-predicate-object. For example, the triple in line 8 of this listing describe the creator of the EMO resource and can be read as *"There is an EMO identified by the URI (rdf:about="oai:pubmedcentral.nih.gov:2820004") whose creator is Kaabia, Noufel"*. Other RDF statements in the listing relate the EMO resource to another RDF resources and define its relation using a predicate. An example of such RDF statements are the attributes defining subjects to categorise the EMO (<dc:subject>) (line 19-23) or the title of an EMO (<dc:title>) (line 15). The RDF resource that is used to describe the title of the EMO is identified by the URI (rdf:about="oai:pubmedcentral.nih.gov:2820004:title") and is explained in the next section.

Listing 6.1: Example EMO

```
1 <rdf:Description rdf:about="oai:pubmedcentral.nih.gov:2820004">
2     <dc:identifier>http://www.ncbi.nlm.nih.gov/pubmed/20148139</dc:identifier>
3     <dc:identifier>http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2820004/
4                                         </dc:identifier>
5     <dc:identifier>http://dx.doi.org/10.1186/1757-1626-3-11</dc:identifier>
6     <dc:creator>Kaabia, Naoufel</dc:creator>
```

```
    <dc:creator>Jazia, Elhem Ben</dc:creator>
    <dc:creator>Khalifa, Mabrouk</dc:creator>
    <dc:creator>Ghannouchi, Neirouz</dc:creator>
    <dc:creator>Bahri, Fethi</dc:creator>
    <dc:creator>Letaief, Amel</dc:creator>
    <dc:creator>Abdelkader, Atef Ben</dc:creator>
    <dc:creator>Braham, Ahlem</dc:creator>
    <dc:date></dc:date>
    <dc:title rdf:resource="oai:pubmedcentral.nih.gov:2820004:title"/>
    <dc:description rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc"/>
    <dc:rights>Copyright 2010 Jazia et al; licensee BioMed Central
    Ltd.</dc:rights>
    <dc:type> Article </dc:type>
    <dc:subject rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term6"/>
    <dc:subject rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term9"/>
    <dc:subject rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term7"/>
    <dc:subject rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term4"/>
    <dc:subject rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term10"/>
    <dc:publisher>BioMed Central</dc:publisher>
  </rdf:Description>
```

6.2.2 The Titles and Description Resource

The Title and Description resources are RDF resources created at the enrichment process of the LEMO system. Both are used to describe the literal text describing the EMO's title or description as harvested from the web. Furthermore, these resources are used to hold relations between the EMO and the annotations discovered during the enrichment process. The resources are identified with URIs invented based on the EMO resource URI they are describing. In the EMO sample (listing 6.1), the EMO with the URI (`rdf:about="oai:pubmedcentral.nih.gov:2820004"`) is related to a Title resource with the URI (`rdf:about="oai:pubmedcentral.nih.gov:2820004:title"`) via the predicate `<dc:title>`. The same case applies for the Description resource that is related via the `<dc:descriptoin>` predicate with the Description resource identified by the URI (`rdf:about="oai:pubmedcentral.nih.gov:2820004:desc"`). The RDF statements describing the Title and the Description resources are detailed in listing 6.2, and list 6.3 respectively.

Listing 6.2: Example EMO's title

```
<rdf:Description rdf:about="oai:pubmedcentral.nih.gov:2820004:title">
   <rdf:value>A case of primary biliary cirrhosis associated with pernicious
   anemia: a case report</rdf:value>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term1"/>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term2"/>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term3"/>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term4"/>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term5"/>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term6"/>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term7"/>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term8"/>
   <lemo:lemoTitleAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:title:term9"/>
 </rdf:Description>
```

Listing 6.3: Example EMO's description

```
<rdf:Description rdf:about="oai:pubmedcentral.nih.gov:2820004:desc">
   <rdf:value>Primary biliary cirrhosis is often associated with autoimmune
   diseases. However, its association with pernicious anemia has rarely been
   reported.</rdf:value>
   <lemo:lemoDescAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term9"/>
   <lemo:lemoDescAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term3"/>
   <lemo:lemoDescAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term2"/>
   <lemo:lemoDescAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term6"/>
   <lemo:lemoDescAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term5"/>
   <lemo:lemoDescAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term11"/>
   <lemo:lemoDescAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term8"/>
   <lemo:lemoDescAnnotation
   rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term7"/>
```

```
11    <lemo:lemoDescAnnotation
      rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term1"/>
12    <lemo:lemoDescAnnotation
      rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term4"/>
13    <lemo:lemoDescAnnotation
      rdf:resource="oai:pubmedcentral.nih.gov:2820004:desc:term10"/>
14  </rdf:Description>
```

The Title and Description resources are described with similar RDF statements. Both have the same functionalities, that is, to describe the literal text value of that resource using the predicate <rdf:value> (line 2 in both listing 6.2, and listing 6.3) and to describe their relations with the Term resources representing its annotations. Different predicates are used to describe the relations between the two resources and their Term related resources. For the Title resource (listing 6.2), the <lemo:lemoTitleAnnotation> (line 3-11) is used. While, for the Description resource (listing 6.3), the predicate used is the <lemo:lemoDescAnnotation> (line 3-13). These predicates were proposed in the LEMO metadata schema to differentiate between the concepts annotating the text in the title or the description of an EMO. These concepts are weighted and filtered to select a smaller set of concepts that define the subject of that EMO. This is conducted in the enrichment process of the LEMO system.

6.2.3 The Term Resource

The Term resources are the object values of many RDF statements describing an EMO. All of the concepts annotated in the title or the description text are represented as Term resources that are related to the Title and Description resources as explained in the previous section. Also, they are used to describe the subject <dc:subject> of an EMO (listing 6.1). The Term resource is described as a collection of RDF statements using predicates proposed in the LEMO metadata schema. The detailed explanation of each predicate and its usage was discussed in the LEMO metadata implementation (Chapter 4, section 4.5.1). An example of a Term re-

source description is provided in listing 6.4 and it is apparent from the resource URI (rdf:about="oai:pubmedcentral.nih.gov:2820004:title:term8) that it is related to the Title resource of an EMO.

Listing 6.4: Example of a term discovered in EMO's title

```
1  <rdf:Description rdf:about="oai:pubmedcentral.nih.gov:2820004:title:term8">
2    <rdf:value>3</rdf:value>
3    <lemo:lemoTermDef></lemo:lemoTermDef>
4    <lemo:lemoTermSynon>Pernicious anemia (disorder);Megaloblastic anemia due to
     impaired absorption of cobalamin;Addison s anemia;Biermers
     anaemia;Megaloblastic anaemia due to impaired absorption of
     cobalamin;Biermer anemia;Addisonian pernicious anaemia;Pernicious anaemia;PA
     - Pernicious anaemia;Addisonian pernicious anemia; </lemo:lemoTermSynon>
5    <lemo:lemoTermClassLabel>Pernicious anemia</lemo:lemoTermClassLabel>
6    <lemo:lemoTermClassID>http://purl.bioontology.org/ontology/SNOMEDCT/84027009
     </lemo:lemoTermClassID>
7    <lemo:lemoTo>69</lemo:lemoTo>
8    <lemo:lemoFrom>53</lemo:lemoFrom>
9    <lemo:LemoTermText>PERNICIOUS ANEMIA</lemo:LemoTermText>
10   <lemo:lemoTermID>http://purl.bioontology.org/ontology/SNOMEDCT/84027009
     </lemo:lemoTermID>
11 </rdf:Description>
```

The RDF statements describing the Term resource provide information about which part of the title text is annotated. The values of the predicates <lemo:lemoFrom> and <lemo:lemoTo> are (from 53 to 69) respectively. These values are the indices of the text annotated in the title of an EMO. If mapped to the literal text value of the <rdf:value> predicate (listing 6.2) (line 2), the text *"Pernicious Anemia"* is extracted. This part of the text is annotated by a concept in the SNOMED CT ontology, and details about that class are retrieved and described as RDF statements of the Term resource. As explained in the LEMO metadata schema implementation, the functionality of the two predicates <lemo:lemoTermID> and <lemo:lemoTermClassID> might be confusing in this research. Both predicates were implemented to enable the use of multiple ontologies for annotating the content of LEMO store. A term can be annotated with several class ontologies via the bio-Portal annotator tool incorporated in the LEMO system. Hence, the same term

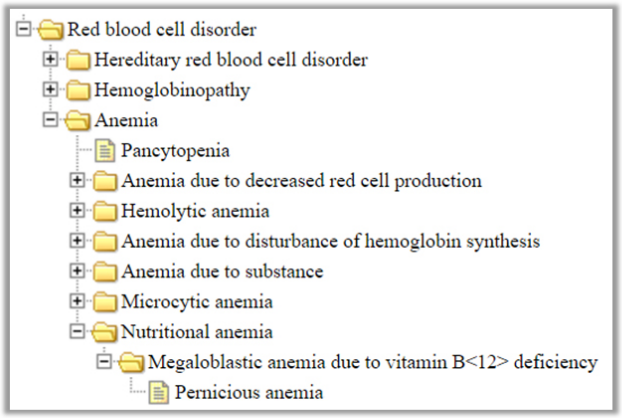
annotated can have multiple classes annotating it. The ontology class annotating any term is described by the predicate <lemo:lemoTermClassID>, and it is a URL value retrieved from the SNOMED CT ontology. The collection of ontology classes URLs are represented as Class resources and their relations are maintained by the RDF statements describing them. These Class resources can be easily referenced from the Term resources using these URLs .

6.2.4 The Class Resource

The URL of an ontology class is used as its URI identifier since the ontology classes exist on the real web and can be dereferenced with this URL. The Term resources are related to these Class resources using their URLs provided in their description. Moreover, two Term resources are similar to each other if they are annotated with the same ontology class. A Class resource is simply described with RDF statements that define the class label and its hierarchical relation to its descendent classes only. Using these RDF statements describing the ontology classes, a tree structure can be derived to represent any ontology class. For example, the ontology class annotating the Term resource (listing 6.4, line 6) is identified by the URL value of the predicate <lemo:lemoTermClassID>. This ontology class is represented as an RDF resource in the Class resource shown at the bottom of Figure 6.5. The figure illustrates this ontology class and other Class resources that are adjacent to it as noted by the arrows. The LEMO system is capable of reading these RDF statements and builds a hierarchical tree of the ontology classes. Four Class resources are illustrated in the figure and build part of the tree shown at the top of Figure 6.5. The Class resource at the bottom has one triple that describes the class label that is *Pernicious Anemia*. It has no other predicates to describe its adjacency with other Class resources making it a leaf node as shown in the tree. The other Class resources are described with RDF statements that define its adjacency to any descendent ontology classes using the predicate <lemo:adjacentTo>. The use of RDF is shown to be highly effective

in this sample. It is easy to read the descendent nodes of one class ontology as it is easy to read the ancestor of one ontology class.

It was intended to keep the Class resource description very simple because these resources are processed and frequently updated as new annotations added to the LEMO dataset. It is considered as a separate RDF model that is not connected to any of the other resources that might affect its processing. Adding a new Term




```
<rdf:Description rdf:about="http://purl.bioontology.org/ontology/SNOMEDCT/271737000">
  <lemo:lemoTermClassLabel>Anemia</lemo:lemoTermClassLabel>
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/14514008</lemo:adjacentTo>
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/61261009</lemo:adjacentTo>
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/424988008</lemo:adjacentTo>
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/127034005</lemo:adjacentTo>
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/59106005</lemo:adjacentTo>
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/66612000</lemo:adjacentTo>
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/234349007</lemo:adjacentTo>
</rdf:Description>

<rdf:Description rdf:about="http://purl.bioontology.org/ontology/SNOMEDCT/66612000">
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/49472006</lemo:adjacentTo>
  <lemo:lemoTermClassLabel>Nutritional anemia</lemo:lemoTermClassLabel>
</rdf:Description>

<rdf:Description rdf:about="http://purl.bioontology.org/ontology/SNOMEDCT/49472006">
  <lemo:adjacentTo>http://purl.bioontology.org/ontology/SNOMEDCT/84027009</lemo:adjacentTo>
  <lemo:lemoTermClassLabel>Megaloblastic anemia due to vitamin B&gt;12&lt; deficiency
                                                    </lemo:lemoTermClassLabel>
</rdf:Description>

<rdf:Description rdf:about="http://purl.bioontology.org/ontology/SNOMEDCT/84027009">
  <lemo:lemoTermClassLabel>Pernicious anemia</lemo:lemoTermClassLabel>
</rdf:Description>
```

Figure 6.5: Class resources and the navigation menu generated

resource that annotates an ontology class that was not stored in the LEMO RDF store requires updating the adjacency list of one or more Class resources. Some ontology classes can have multiple parents. Hence, the relations between the Class resource results in having an RDF graph that can be used for building linkages.

6.3 Formal Description of the RDF store

In this section, a formal representation of the RDF resources composing the LEMO RDF store is detailed. The following notations are used to denote the RDF store components. Defining these notations will be beneficial for detailing the content of the LEMO store and explaining the techniques developed for evaluating its content. The different RDF resources explained in the previous section are given the following notations. The RDF store consists of EMOs described as RDF resources denoted as the set EMO where the size of this set is $|EMO| = i$ such that:

$$EMO = \{emo_1, emo_2, emo_3, \cdots, emo_i\} \tag{6.1}$$

Moreover, all the Term resources related to the Title resources in the RDF store are denoted as the set KT, and all the Term resources related to the Description resources in the RDF store are denoted as KD , while the overall collection of Term resources is dented as K such that:

$$K = KT \cup KD \tag{6.2}$$

A subset of these Term resources are related to the EMO resource itself and used to describe the subject of that EMO. Hence, the Term resources related to the subject of the EMO are denoted as KS where $KS \subseteq K$ since they might be selected from the Title related terms or the Description related terms.

Based on the previous notations, the following notations can be used to de-

note each EMO resource. For each emo_i, the collection of Term resources related to its Title resource is denoted by KT_i, while the collection of Term resources related to its Description resource is denoted by KD_i. Therefore, an EMO can be represented by a vector of Term resources that annotates its metadata such that $emo_i = K_i$. Where:

$$K_i = KT_i \cup KD_i \tag{6.3}$$

where $KT_i = \{kt_{i_1}, kt_{i_2}, kt_{i_3}, \cdots, kt_{i_n}\}$, and $KD_i = \{kd_{i_1}, kd_{i_2}, kd_{i_3}, \cdots, kd_{i_m}\}$. The sets representing the Term resources in the RDF resources are the result of union relationships between all the Term resources related to each EMO, such that:

$$K = K_1 \cup K_2 \cup \cdots \cup K_i \tag{6.4}$$

$$KT = KT_1 \cup KT_2 \cup \cdots \cup KT_i \tag{6.5}$$

and,

$$KD = KD_1 \cup KD_2 \cup \cdots \cup KD_i \tag{6.6}$$

Moreover, the set of Term resources selected to represent the subject `<dc:subject>` of an EMO emo_i are denoted as KS_i where $KS_i \subseteq K_i$ and can be defined as $KS_i = \{ks_{i_1}, ks_{i_2}, ks_{i_3}, \cdots, ks_{i_p}\}$. These Terms can be belong to either KT_i or KD_i. Also, the collection of all Term resources selected to represent subjects of all the EMOs KS is defined as $KS = KS_1 \cup KS_2 \cup \cdots \cup KS_i$.

The goal of annotating EMOs with concepts from ontologies is to build relations between these EMOs. Originally, the ontologies can be represented as taxonomic hierarchical relations of concepts. Each ontology class might have one or more parent. Hence, the ontology can be referred to as a graph of vertices and edges. In the LEMO system, the SNOMED CT ontology is used to enrich the LEMO dataset.

The relations between its concepts are used to build the linkages between the EMOs in the dataset. In order to understand how the ontology taxonomies are used to build relations in the LEMO dataset, a formal representation of these relations is necessary.

Since, the SNOMED CT ontology has been incorporated in the LEMO system for enriching the final LEMO dataset, then the taxonomy of this ontology is organised in a graph structure denoted by:

$$G_{snomed} = (V_{snomed}, E_{snomed}) \tag{6.7}$$

The G_{snomed} graph is not physically stored in the LEMO system. It is used as a reference for building relations between the all the Term resources K of EMO to build a smaller sub-graph of G_{snomed} referred to as G_{emo} where:

$$G_{emo} = (V_{emo}, E_{emo}) \tag{6.8}$$

All of the Terms discovered in LEMO and stored in set K reference a class that is a vertex in G_{snomed}. Connecting the set of classes stored in K results in the G_{emo} graph, where classes of $K \subseteq V_{emo}$ and other vertices are transition vertices which are not annotated in the LEMO dataset but used to build the graph edges. This graph G_{emo} is stored in the LEMO RDF store and represents its bottom layer (figure 6.4), and is used as the basis for processing the relations between EMOs at the top layers.

Since the set K consists of a smaller sets of K_i for each emo_i, then based on the G_{emo} graph, each EMO can be presented as a graph of related Term resources referred to as G_{emo_i} where:

$$G_{emo_i} = (V_{emo_i}, E_{emo_i}) \tag{6.9}$$

Since all the keywords K_i discovered in each EMO emo_i reference a class

in the SNOMED CT ontology, then $K_i \subseteq V_{emo_i}$ and other vertices are transition nodes needed for building hierarchical relations between the terms discovered in each EMO. Formally representing the relations between the RDF resources in the graph structure is important for understanding how the linkages in the LEMO store are created. Equation 6.10 is valid for each emo_i and represents the graphs dependencies in the LEMO system.

$$G_{emo_i} \subseteq G_{emo} \subseteq G_{snomed} \tag{6.10}$$

The graph representation of each EMO in the LEMO store has been used for weighting the importance of the Term resources related to each EMO. The subject selection process performed during the enrichment process (Chapter 5, section 5.3) depended on the graph structure of each EMO represented as G_{emo_i} in order to weight and rank its Term resources.

6.4 The RDF Store Content

The final RDF store content resulting from running several experiments for aggregating and integrating web data sources is presented in this section. It details the number of EMOs harvested, the annotations added to enrich their metadata, and the number of links that can be generated in the dataset. Details about the SNOMED CT ontology and the number of concepts used in annotating the LEMO dataset are detailed. Furthermore, the results of categorising the EMOs into subjects are explained in addition to the detailed number of links that can be generated between the EMOs in the RDF store.

6.4.1 The Dataset

The RDF store is composed of EMO resources, Title resources, Description resources, Term resources, and Class resources. The type of an EMO resource can

Table 6.1: The final LEMO dataset components

EMO type	$\|EMOs\|$	$\|KT\|$	$\|KD\|$	$\|K\|$
Article	8742	56708	307431	364139
Video	1259	3297	5348	8645
Blog	461	1494	9766	11260
Total	10462	61499	322545	384044

be either an article, a blog, or a video, and that is described using the predicate `<dc:type>` depending on the web data source that is hosting this EMO. Each EMO resource is related to one Title resource and one Description resource via the predicates `<dc:title>` and `<dc:description>` respectively. Each Title or Description resource is related to Term resources via the predicates `<lemo:lemoTitleAnnotation>` or `<lemo:lemoDescAnnotation>`. Each Term resource references an ontology concept identified by a class URL stated as the value of the predicate `<lemo:lemoTermClassID>`. Each ontology concept is defined as a Class resource that is related to the collection of ontology concepts annotating the LEMO dataset. Many Term resources might reference the same ontology class. In other words, EMOs discussing similar topics might use the same wording in their titles or descriptions, therefore, these words can be annotated with the same concept of an ontology. The number of EMOs grouped by type and the number of annotations discovered in the title or description text are detailed in Table 6.1.

The majority of the dataset content is harvested from the *PubMed library* using OAI-PMH harvesting endpoint. Also, videos and blogs were harvested from a list of *YouTube* channels and blogging platforms managed by well-known medical institutes using an RSS feeds reader endpoint. The metadata provided for videos and blogs are not well documented since they are user generated content. For example, the description field of a video, published on *YouTube*, might be missing, and the title field is not descriptive enough for the content of the video. Thus, the annotations resulted from enriching videos and blogs are less than the articles annotations, as shown in Table 6.1. In the LEMO dataset, EMOs of type video are

the least enriched EMOs in the dataset compared to their size.

6.4.2 The Ontology

The ontology incorporated in the LEMO system for enriching the EMOs stored in the RDF store is the SNOMED CT ontology. It is an acronym for Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT). The ontology provides a comprehensive healthcare terminology that contains interrelated concepts, supported by synonyms and definitions [Stearns et al., 2001]. The ontology classes used to represent the Term resources annotating the EMOs are represented as Class resources. The collection of Class resources and their relations are denoted by the graph G_{emo} that represent the bottom layer of the RDF store (figure 6.4). Hence, the G_{emo} graph is considered a subset of the G_{snomed} graph (equation 6.10). Details about the number of the concepts in each graph and the depth of that graph are detailed in Table 6.2. It presents a comparison between the SNOMED CT ontology and the sub-graph resulted from using its concepts for enriching the LEMO dataset.

The number of the ontology concepts that are stored as Class resources in the RDF store is a small subset of the SNOMED CT classes as detailed in the table. The number of EMOs described in the LEMO RDF store is more than 10,000 which are annotated with more than 29,000 concepts from the SNOMED CT ontology. Harvesting more data from the web and enriching it might increase the number of ontology concepts used and increase the size of the subgraph G_{emo} representing the relations between the ontology concepts. The maximum depth represents the length of the deepest branch in the SNOMED CT taxonomy. It is organised based on the is-a hierarchical relations between its concepts. The maximum depth of the G_{emo}

Table 6.2: Comparison between the SNOMED CT and LEMO graph

Metrics	G_{snomed}	G_{emo}
Number of Classes	316031	29283
Maximum Depth	28	25

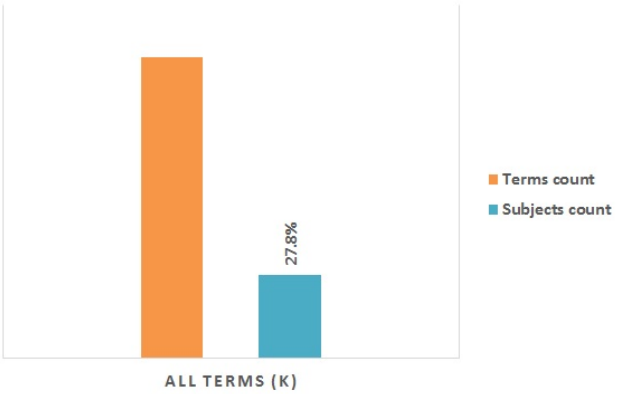
graph is 25 levels. That means that some concepts annotating the EMOs are in the lower levels of the ontology hierarchy.

6.4.3 Enriched Dataset

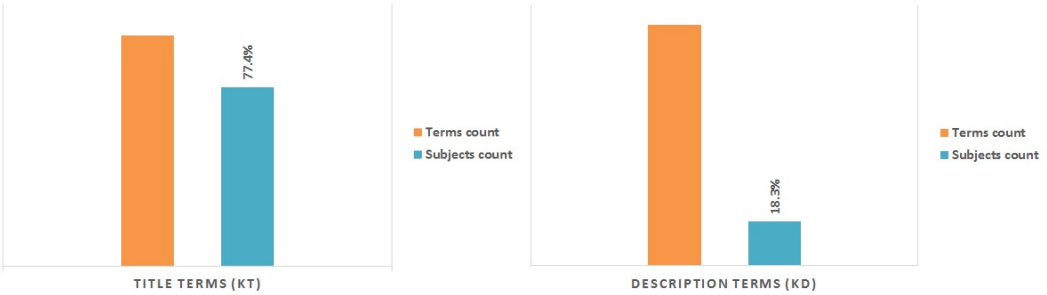
In the enrichment process of the LEMO system, the Term resources related to each EMO are weighted and filtered to choose a smaller set of Term resources that will be referenced by the predicate `<dc:subject>` describing an EMO resource. For each EMO, the set of Term resources relating to it can be represented as a graph structure based on the relations between the ontology classes referenced in these Terms. The subject selection process is dependent on the weight values assigned to each Term resource in the graphs derived from each EMOs. The results of the subject selection process are illustrated in Figure 6.6. The Term resources related to the subject attribute of an EMO are denoted by KS. Hence, from the set of Term resources (K) composing the RDF store, 27% were selected to describe the `<dc:subject>` of an EMO as illustrated in figure 6.6a. This subset is either selected from the Term resources related to the Title of EMOs (KT) or the Term resources related to the Description of an EMO (KD). A percentage of 77% of the Terms related to the Title of an EMOs were selected to describe the subject attribute in the RDF store (figure 6.6b), compared to only 18% of the Term resources related to the Description of an EMO (figure 6.6c). The majority of these Term resource were terms annotated in the Title on an EMO. Hence, the results are compliant with the general practice that the subject or the topic of an EMO can be deduced from its title since it is more general and less detailed.

6.4.4 Links Analysis

After the subject selection process, links are generated between EMOs based on their categorising Terms (KS). A link exists between two EMOs if they have at least one similar annotated class in their list of annotations represented as Term



(a) All Terms



(b) Terms annotated in the Title Resources (c) Terms annotated in the Description Resources

Figure 6.6: Subject selection results

resources related to the Title or the Description of an EMO. A link exists between emo_i and emo_j if the intersection between the two sets of Term resources described in the predicate `<dc:subject>` is equal to one or more ontology classes, that is $|KS_i \cap KS_j| >= 1$. Additionally, links can be generated based on the similarity of Terms annotated in the Title resources of EMOs, or the Terms annotated in the Description resources of EMOs, that is $|KT_i \cap KT_j| >= 1$ and $|KD_i \cap KD_j| >= 1$ respectively. The links count between two EMOs is the number of common ontology classes annotated in their Terms. Also, the links in the LEMO store are considered direct links, therefore, if there is a link from node a to node b the link will be counted twice instead of once. Having a large number of Term resources annotating the EMOs, several links can be generated based on these Terms whether they are referenced by the Subject `<dc:subject>`, the Title annotation `<lemo:lemoTtileAnnotation>`, or

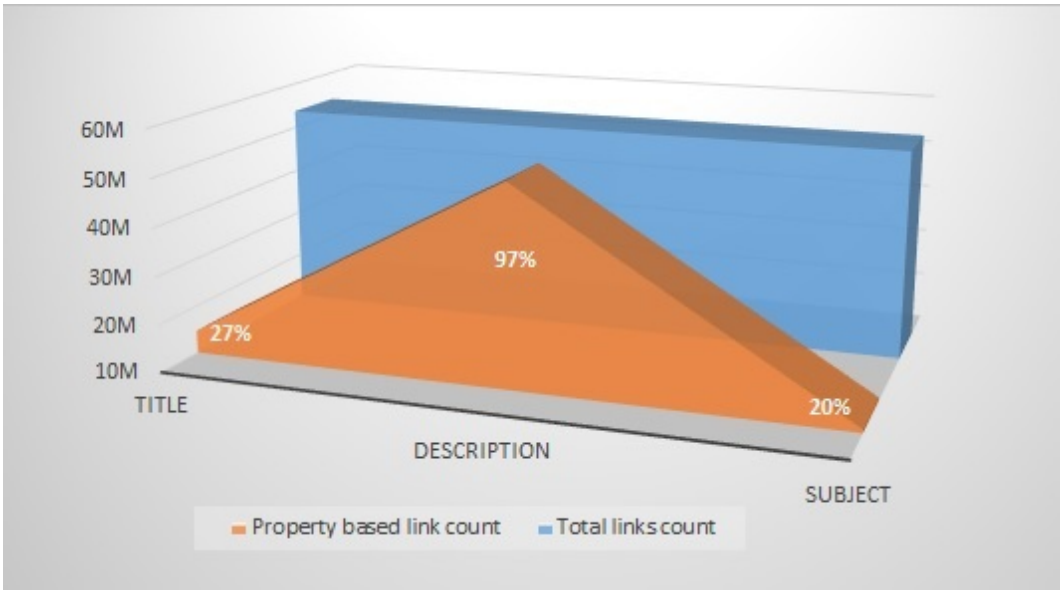


Figure 6.7: Percentages of links made based on the properties of the EMOs

the Description annotations `<lemo:lemoDescAnnotation>`. Assuming that the Term resources referenced by the Subject attribute are more representative of the EMOs content, these Subject Terms (KS) are considered the basis of the linkage process in the LEMO system. Figure 6.7 illustrates the number of links generated in the LEMO store based on all the Terms discovered with percentages detailing the links generated based on different metadata elements: Title Terms (KT), Description Terms (KD), and Subject Terms (KS).

The total number of links made based on all the Terms annotating the EMOs (K) is large (more than 50 million). Terms referenced by the Title or the Description resources of EMOs can link EMOs to any other EMO annotated related to a Term resource that map to the same ontology class. Hence, most of the links are generated based on the Terms annotating the Description resource of an EMO since there are greater in number as detailed in Table 6.1. The least number of links was generated based on the Terms referenced by the Subject of EMOs. These links are generated between two EMOs if they have at least one ontology class in their lists of Term resources referenced by the Subject attribute of these EMOs. The links based on

the Subject Terms are 20% of the overall number of links (around 10 million). The results indicate that the average number of links generated for each EMO in the LEMO store is around 100. Thus, the number of links for each EMO varies based on its related number of Terms. The quality of the links based on the Terms defining the subject of an EMO is stronger than those based on the Title or the Description related Terms. The technique followed in the subject selection process aims to choose only the repeated concepts annotated in the EMOs and it considers the relation of the concepts with other related concepts annotated in the EMO.

6.5 Summary

The LEMO system has been implemented to harvest, enrich, and build one coherent dataset of EMOs of different types collected from the web. This chapter details the final result of running several experiments in LEMO system resulting in the LEMO dataset. The EMOs has been harvested from Web 2.0 websites such as *YouTube* and bloggin platforms using the RSS feeds harvesting endpoint developed in the LEMO system. Also, large number of articles were collected from the *PubMed Library* using the OAI-PMH harvesting endpoints incorporated in the LEMO system. The final dataset of harvested EMOs is more than 10,000 EMOs of different types concerned with various medical topics. The process of harvesting the data was not supervised and was performed by the LEMO system automatically. Hence, the topics covered in the EMOs varies and not predefined. The LEMO system has enriched the EMOs metadata with concepts of the SNOMED CT ontology by annotating its metadata using BioPortal annotator API. The final dataset resulted from the LEMO system is the LEMO dataset. It was stored in an RDF store that has been detailed in this chapter. The final components of the LEMO dataset has been explained in this chapter in order to understand its structure which is necessary for understanding the consequent chapters.

In conclusion, this chapter has addressed the research objective **O8**: "Describe the RDF store that is managed by the LEMO system for organising the EMOs metadata represented in the LEMO metadata schema". The outcome of this chapter is considered an introduction for evaluating the content of the LEMO dataset and the techniques developed for accessing its content in the following chapters.

Chapter 7

Information Retrieval by Browsing

7.1 Introduction

Information seeking is considered the main task for human interaction with the web. The methods developed for searching and exploring information from any organised data source to satisfy users' needs are referred to as information retrieval methods [Zhang, 2008*a*]. In any information retrieval system, the process of information organisation is invisible to users but its indispensable for setting out the methods of accessing and retrieving data. In this research, the organisation of the EMOs in the RDF store detailed in the previous chapter (chapter 6) defines the methods developed for accessing and retrieving its content. The RDF store content has been built by aggregating EMOs from distributed web data sources and mapping its metadata into the LEMO metadata schema which has been enriched with the SNOMED CT ontology. The technique developed for navigating and browsing the RDF store utilises the enrichments added to the EMOs metadata based on the SNOMED CT ontology and uses it to browse the RDF store. While the following chapter (chapter 8) introduces an ontology-based query searching technique based

on the ontology concepts as well. Both browsing and querying retrieval techniques were implemented as a web access interface that is tested by conducting experiments that simulate users' access. As the RDF store is not published for public access, the web interface was implemented in a local server. Experiments were conducted to test this web interface by simulating users' access to the RDF store and so evaluate the final LEMO dataset built.

The linkages between the EMOs stored in the RDF store were built based on the added Term resources annotating its titles and descriptions with ontology concepts. In order to evaluate these linkages, the browsing method developed in the web browsing interface was based on the ontology classes and referred to as the ontology-based browsing method. This method introduces a user interface with a navigational menu that indicates the organisation of the RDF store content in categorised concepts. Users' interaction with this menu triggers the process of retrieving the information needed from the RDF store. Since the LEMO dataset was annotated with concepts from the SNOMED CT ontology, these concepts act as text surrogates of the EMOs metadata and are used for retrieving the EMOs. Generally, browsing is considered an interactive search activity in which the search actions are controlled by the users' information needs [Büttcher et al., 2010]. Therefore, the ontology-based browsing method proposed in this chapter adheres to the general properties of any browsing behaviour in terms of search results relevance, continuity, and granularity [Zhang, 2008*b*]. The results of browsing the RDF store are retrieved as an unordered set of EMOs that are evaluated to measure its coherence according to the linkages created by the LEMO system.

7.1.1 Chapter Objectives

For the sake of evaluating the coherence of the RDF store and validating the use of the SNOMED CT ontology for organising and categorising its content, this chapter aims to present an ontology-based browsing method, its implementation, and the

required experiments for evaluating its results. It addresses the research objective **O9**: "Develop ontology-based method for browsing the LEMO dataset resulted from the LEMO system and evaluate the similarity between the results retrieved while browsing". Users browsing behaviour was simulated in the experiments run to demonstrate the results of ontology-based browsing method over the RDF store. The retrieved results are unordered sets of EMOs that provide an overview of the LEMO triple store content. Introducing other techniques are necessary for validating the results of the browsing method developed. Hence, clustering techniques have been incorporated in validating the linkages between the EMOs in the LEMO store, which helps in discovering strongly connected communities of EMOs inside larger and complex networks of EMOs. The simulation of the browsing behaviour using the proposed method for navigating the content of the RDF store is explained in the following scenario.

7.1.2 Scenario Example

The browsing scenario for a user navigating through the menu in Figure 7.1 is given as an example to explain the ontology-based browsing behaviour.

Scenario:

A student is searching the RDF store for EMOs related to the topic "Inner ear structure". Clicking on this node triggers a query for retrieving all EMOs annotated with this topic or any of its descendant sub-topics. The hierarchical structure of the ear part shown in Figure 7.1 details the parts composing the "Inner ear structure" class as defined in the SNOMED CT ontology. The student browsing this topic retrieves EMOs relevant to any part of the inner ear structure such as the "semicircular canal" and the "Vestibular". EMOs annotated with the node selected or any of its descendent nodes are retrieved.

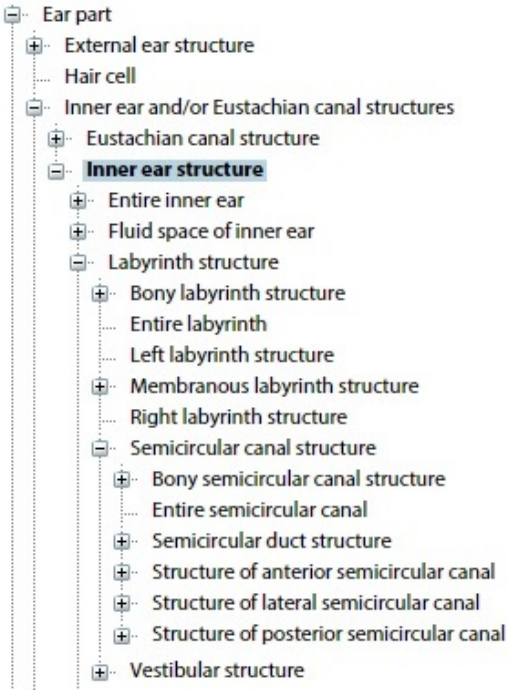


Figure 7.1: Ear part from SNOMED CT

The menu shown in Figure 7.1 is a screen shot of the SNOMED CT ontology as displayed in the BioPortal repository[1]. The navigational menu developed in the browsing web interface was based on the collection of Class resources used to annotate the Term resources in the LEMO RDF store. The Class resources and their relations stored in the LEMO RDF store are considered a sub-graph of the SNOMED CT ontology graph.

7.1.3 Chapter Outline

This chapter is outlined as follows. Firstly, the development of the navigational menu interface for the LEMO RDF store is explained. Then, the ontology-based browsing method is detailed along with the evaluation criteria for its usage. Also, the overview results of exploring the LEMO store are discussed. Finally, methods for clustering the browsing results are explained and discussed for validating the

[1]http://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=root

results of the ontology-based browsing method.

7.2 Ontology-based Navigation

Browsing is an important means to glance over the content of any dataset. A well-organised information space assures successful browsing [Zhang, 2008*b*]. The LEMO dataset has been built by aggregating content of diverse web data sources. Hence, an organised dataset is not guaranteed as no human involvement is considered.

Hierarchical structures are widely used information organisation methods for browsing. Since the LEMO dataset is annotated using ontology classes, the use of ontology hierarchical relations can serve as the navigational guidance of the LEMO dataset. Ontologies define concepts and relations between these concepts. The EMOs have been annotated with concepts from the SNOMED CT ontology. A sample of the hierarchical tree representing the SNOMED CT ontology, its concepts and parts of its relations is illustrated in Figure 7.2. These annotations were repre-

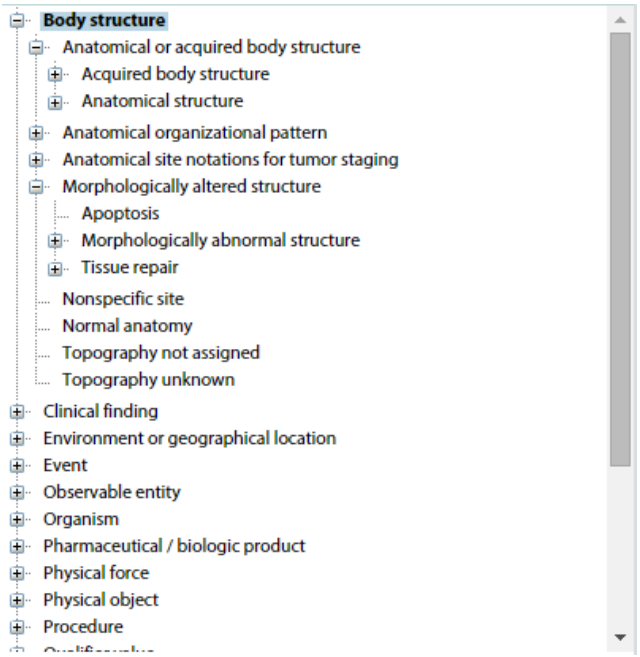


Figure 7.2: SNOMED CT hierarchical tree sample

sented as Term resources stored in the LEMO RDF store. The Term resources have been used to describe the subject elements in the LEMO metadata schema. Hence, the ontology classes are the categories classifying the EMOs in the RDF store. In the ontology-based browsing method proposed, the set of ontology classes used to describe the Term resources categorising the EMOs, and the relations between these classes, have been used to represent a hierarchical structure for navigating the RDF store. A collection of ontology classes are described and organised in the RDF store as it has been explained in the previous chapter (Chapter 6). It forms a subset of the SNOMED CT ontology graph G_{snomed} and is denoted by G_{emo} that is referred to as a LEMO graph. This graph was illustrated as the bottom layer of the RDF store. It presents the Class resources and their relations as retrieved from the SNOMED CT ontology. This graph is considered the basis for browsing the RDF store after it is converted into a tree-like view as shown in Figure 7.3. The figure

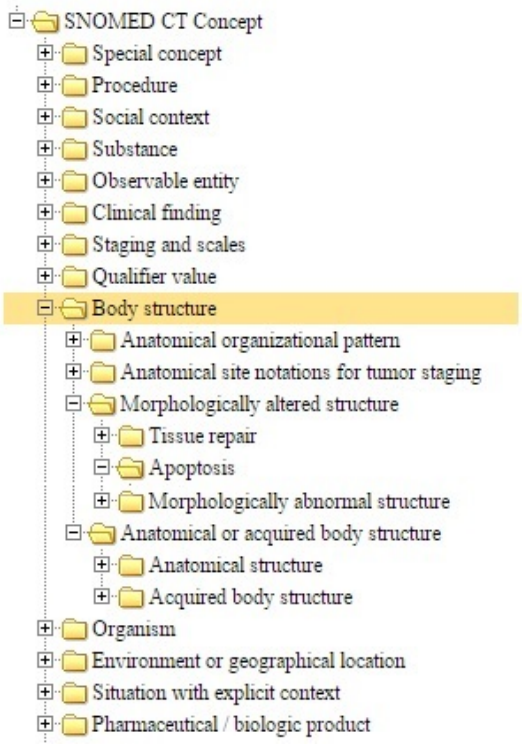
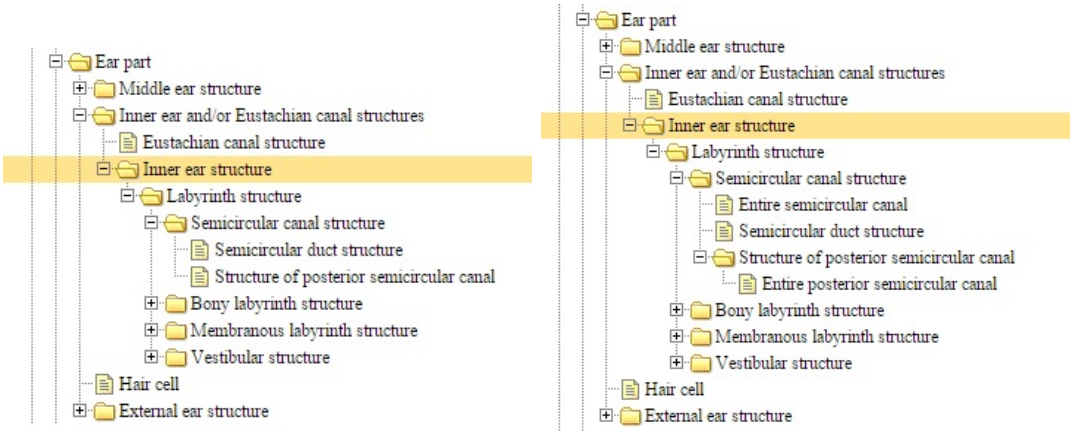


Figure 7.3: LEMO navigation menu

illustrates the same node in the SNOMED CT as it is represented in figure 7.2. It is noticeable by comparing the two figures that the LEMO graph represents a subset of the SNOMED CT ontology.

The LEMO system was developed to build the RDF store that aggregates EMOs from diverse web data sources using harvesting endpoints. Therefore, EMOs can be added to the RDF store, and further enrichments of its metadata are performed. Consequently, SNOMED CT ontology concepts might be added to the RDF store, thus updating the LEMO graph nodes and, as a result, updating the LEMO navigational menu. For example, Figure 7.4a illustrates a branch from the LEMO navigational menu that describes the same node presented in Figure 7.1 from the SNOMED CT ontology. Adding new EMO resources to the RDF Store and enriching these EMOs can result in having further Class resources added to the RDF store. The update of the Class resources description to represent its relations can be reflected directly on the LEMO graph used to build the LEMO navigational menu as shown in Figure 7.4b. Two more nodes are added to this figure as a result of new EMOs being described and stored in the LEMO RDF store. Hence, the ontology-based navigation is considered the method of indexing and navigating the



(a) LEMO navigation menu (b) Updated LEMO navigation menu

Figure 7.4: Snippet of LEMO navigational menu

RDF store as it can easily reflect the topics covered in the LEMO dataset.

7.3 Ontology-based Browsing

The navigational menu presented in Figure 7.3 illustrates tree-like visualization for the LEMO graph G_{emo}. Some of the vertices of this graph are ontology classes representing Term resources annotating the EMOs. These Term resources act as text surrogates of the EMOs metadata. The Term resources related to an EMO emo_i are denoted as K_i where the collection Term resources can either be Terms annotating the Title resources or Terms annotating the Description resources as $K_i = KT_i \cup KD_i$. A subset of these K_i is selected to represent the subject element of an EMO denoted as KS_i. The formal representation of the RDF store components has been detailed in the previous chapter (chapter 6, section 6.3). Now that an overview is given about the Term resources representing an EMO resource, the Term resources referenced as the subject of EMOs KS are used as the basis of this ontology-based browsing method. The set of KS represents Term resources that were filtered and selected to categorise the EMOs. The process of selecting subjects of EMOs has been explained in the LEMO system implementation (chapter 5, section 5.3).

7.3.1 Methodology

The process flow of the proposed ontology-based browsing method is detailed in figure 7.5. It provides a general overview of the EMOs retrieval process implemented and it interacts with the user interface and the RDF store. The process flow starts with the user interacting with the interface that presents the navigational menu. The user clicks on a concept which in return retrieves the ontology class c' that is considered the input for the ontology-based browsing method. The retrieval process starts with building a query vector by retrieving and weighting the adjacent ontology

classes to c' based on the LEMO graph G_{emo}. The query vector q' is then matched with each emo_i based on its Term resources that represent its subject KS_i. Then, the results are returned and displayed to the user.

The proposed ontology-based browsing method is applied using the Boolean model when matching the query vector to the EMOs in the RDF store. The Boolean model is one of the primary information retrieval models. It is often referred to as the "exact matching" model [Liu, 2011]. The standard Boolean model is widely adopted by a large number of current information retrieval systems. It is also easy to implement and computationally efficient [Ceri et al., 2013*a*]. The formal representation of the LEMO triple store (chapter 6, section 6.3) represent the components of the LEMO RDF store in a formal language which is essential for understanding the rest of this chapter. Based on the boolean model the ontology-based browsing algorithm is detailed in Algorithm 3 for explaining the process of selecting concepts and retrieving relevant EMOs. When users click and choose a concept to browse, then they are declaring their interest in a topic that includes

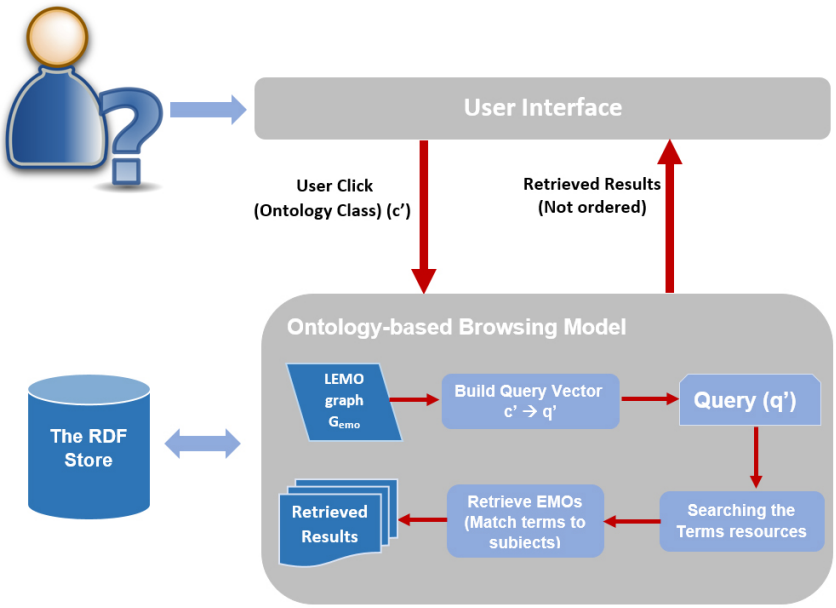


Figure 7.5: Ontology-based browsing model

Algorithm 3 Ontology-based browsing

```
1: Input : ontology class query c' and G_emo
2: Output : set of EMOs qResults
3: procedure BuildQuery(c')
4:     G_c ← sub-tree rooted at c' from G_emo
5:     q' ← getDescendants(c', G_c)
6:     q' = {c_1, c_2, c_3, ··· , c_n}
7:     |q'| depends on the size of G_c
8: end procedure
9:
10: procedure Ontology-based Matching(q')
11:     for each emo_i ∈ EMO do
12:         Read KS_i
13:         if |KS_i ∩ q'| > 1 then
14:             qResults ← emo_i
15:         end if
16:     end for
17:
18:     return qResults
19: end procedure
```

the concept selected and all the descendent concepts from it. In the ontology-based browsing method, the ontology class selected c' is extended to include all the sub-nodes descendent from that class as stored in the G_{emo} to build a query vector named q'. After building a query vector, the next step is to match the EMOs with the query vector. Based on the Boolean model for information retrieval, the ontology-based browsing method performs exact matching with the EMOs in the RDF stores and retrieves any EMO emo_i that is annotated with one or more classes of the query vector q' in the set of Term resources representing that EMO subject KS_i. The EMOs retrieved are stored in a set named *qResults* and returned to the user. The results of this method are not ordered according to its relevance to the topic, although the process of ordering them can be implemented based on the weights of the Term resources related to EMOs. The goal of this ontology-based browsing method is to glance over the LEMO dataset and evaluate the coherence of its EMOs based on the linkages built in the LEMO system. Therefore, the

user feedback about the relevance of the results is not considered in this research. Moreover, the dataset is not published on the web. Hence, the web interface is not deployed on a web server. It is implemented locally for testing the method proposed.

7.3.2 Evaluation criteria

As mentioned in the introduction of this chapter, the proposed ontology-based browsing method adheres to the general properties of any browsing behaviour regarding search results continuity and granularity [Zhang, 2008*b*]. Therefore, evaluating the results of browsing the LEMO RDF store is necessary. In this section, evaluation criteria are introduced to understand the browsing evaluation.

Since the LEMO dataset is large, it is hard to validate its content using experts' judgements. The dataset consists of more than 10,000 EMOs with more than 10 million links connecting these EMOs, as detailed in the previous chapter (chapter 6, section 6.4). Therefore, browsing actions mimicking users' behaviour when accessing the RDF store using the LEMO navigational menu were simulated. The results of these simulations provide an overview of the RDF store content and the linkages built between its EMOs.

The first level of nodes in the navigational menu of LEMO consists of 19 nodes resulted after converting the LEMO graph G_{emo} into a tree-like view. The simulation of browsing these 19 nodes and its descendent classes gives a glance about what topics are covered in the LEMO dataset. As has been clarified in this research, the LEMO dataset has been aggregated from diverse web data sources without human involvement. Hence, the results of navigating the concepts in the menu can indicate the distribution of the EMOs over the ontology concepts that are used to categorise the EMOs. Furthermore, the EMOs retrieved are interlinked based on their annotations discovered in the title or the description of the EMOs. Hence, the browsing results can be represented as a collection of interlinked EMOs forming a graph structure denoted by $G = (V, E)$ where V is the set of EMOs

retrieved, and the E are the links interlinking its vertices. The more the graph vertices are interlinked, the more the EMOs are relevant to each other. Therefore, the link density of these graphs representing the results of browsing the RDF store is calculated based on equation 7.1.

$$Density(G) = \frac{|E|}{|V|.(|V|-1)} \tag{7.1}$$

The density $Density(G)$ measures the links density between the retrieved graph of EMOs where $Density(G) \in [0, 1]$. The larger the density is, the more related the EMOs retrieved are [Lieberam-Schmidt, 2010]. The graph G is updated while browsing the RDF store. It holds the results of each simulation experiment conducted. The simulations start with visiting all the nodes of LEMO graph at one level and continue to navigate all the levels of each node. The maximum level reached for each node is named the *length* of that node. The next section provides the results of conducting the simulation experiments for testing the ontology-based browsing method.

7.3.3 Experimental Results

In order to get an overview of the LEMO dataset distribution over the SNOMED CT concepts, the link density was calculated for all the browsing results retrieved from the first level nodes in the LEMO navigational menu, and comprise 19 nodes. Figure 7.6 illustrates the link density values and the length of each node visited in the experiment that simulates user browsing actions. The figure indicates a positive correlation between the length of the node and the value of the link density in the results of browsing each node. The length of the node represents the number of levels of descendant concepts the node has, and that in return means a higher number of ontology classes descendent from the first level ontology class. The link density results have the tendency to increase when the node length increase. The relation

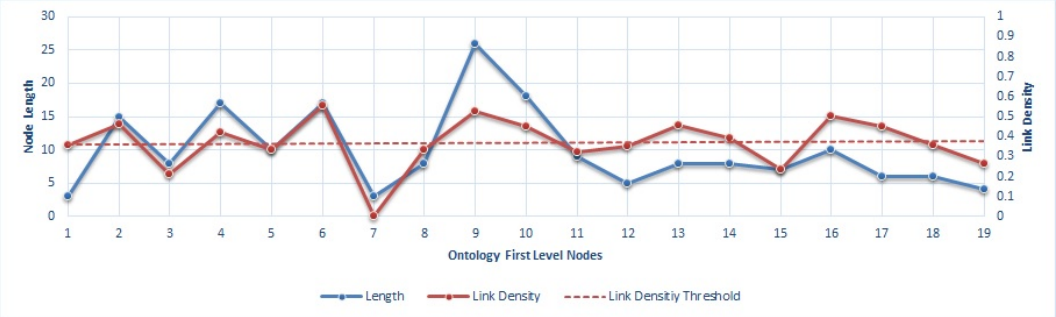


Figure 7.6: The links density vs. node length for first level nodes navigation

between the two variables is measured using the Pearson's correlation coefficient statistical measure [Benesty et al., 2009]. The value of this coefficient for the values presented in Figure 7.6 is 0.633. This value indicates a moderate positive correlation between the length and the density values.

The results of browsing the first level nodes in the LEMO navigation menu gives an overview of the RDF store content. Some nodes have low link density values which indicate a weakly connected set of EMOs retrieved. A threshold is calculated to represent the average link density value discovered by this experiment. The threshold is illustrated in Figure 7.6 in a dashed line. Browsing all the nodes above the threshold have had more dense graphs that indicate retrieving strongly connected EMOs. Hence, further experiments were conducted to browse these dense nodes at deeper levels to evaluate the browsing method proposed. From this point forward, all the evaluation experiments will consider the nine nodes that are above the threshold out of the 19 nodes that were involved in this first experiment.

As for the second experiment conducted for evaluating the browsing behaviour simulated in this chapter, it studies the granularity of the results retrieved when browsing deeper levels of the LEMO navigational menu. In any well-established information retrieval system, the deeper the users browse into the navigation levels, the results decrease in size and increase in density. Therefore, the size and the density of the browsing results are calculated for browsing all the nodes from top

levels in the navigational tree to deeper levels. The average size of the sets of EMOs retrieved and the average link density of all the results at each level are illustrated in Figure 7.7. The nodes in the first levels are the 19 nodes descended directly from the SNOMED CT concept. The experiments span four levels to indicate the significant changes in the browsing results retrieved. The results presented in Figure 7.7 support the general behaviour of any IR system relating to the granularity of the results retrieved. At higher levels of browsing, the size of the results retrieved is large and its content items might not be related to each other. As the users start navigating into deeper levels of the navigational menu, the ontology concepts become more detailed. Thus, the browsing results decrease in size, and only the related EMOs are retrieved resulting in higher link density values.

At this stage of experimenting with the browsing method proposed, the results are promising when it comes to general properties such as relevance and granularity. For further evaluation of the browsing results, clustering techniques are incorporated for detecting communities and sub-communities within the RDF store according to their automatic categorisation with SNOMED CT concepts. However, clustering must be conducted on the dense datasets and will not be efficient if applied on the sparse ones. Therefore, the following experiment browsed the densest

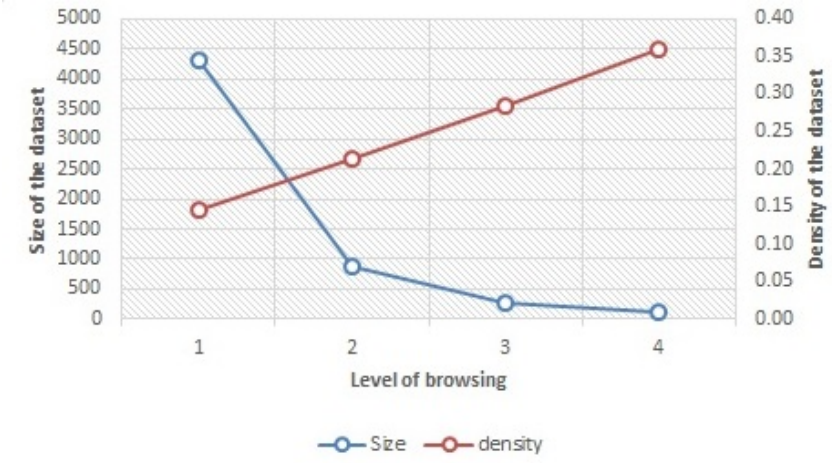


Figure 7.7: Results of browsing different levels

branches descended from the nine nodes that have high link density based on the first experiment results detailed in Figure 7.6. The results of this experiment discover the densest concepts to browse in the LEMO navigational menu. Figure 7.8 illustrates the density of the results retrieved when deeper levels of the nine nodes were browsed. The nine nodes represent the concepts of SNOMED CT ontology as detailed in the figure.

The graph illustrates the link density values of the results retrieved when browsing each node at different levels of depth from 1 to 7. From this experiment, it can be noticed that, for most of the nodes, the link density values start to decrease after level 5. This degradation in the link density might be due to having a smaller number of EMOs retrieved at those levels which are not highly connected to each other. The results of browsing some branches such as "clinical findings" and "body structure" nodes have link density values higher than 0.5 at level 7. Such values indicate that the collection of EMOs retrieved at that level for such nodes are more connected and related to each other. It also indicates that large numbers of EMOs are annotated with the classes descended from these nodes since browsing this deep into the menu is still possible and results in having strongly connected communities retrieved. The results from browsing these two nodes, "clinical findings" and "body structure", are considered for detecting community techniques explained in the next section.

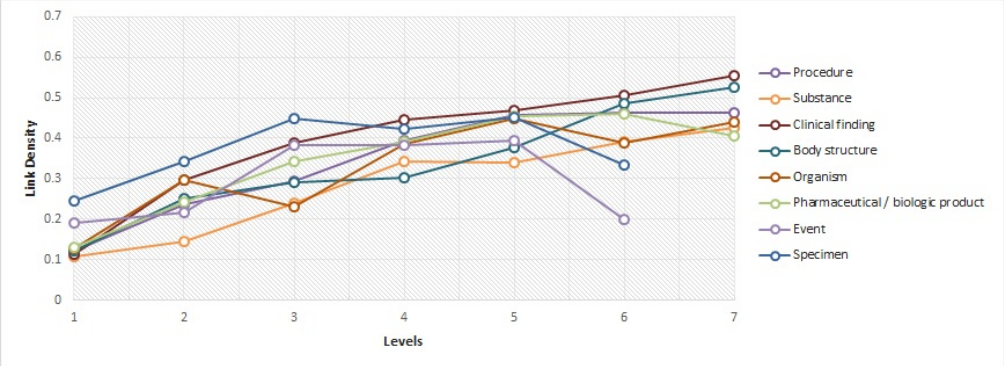


Figure 7.8: The link density score variation at different levels of browsing

7.4 Clustering for Validating Browsing Results

In information retrieval, clustering is used to enhance the results of browsing by grouping (*clustering*) the similar objects retrieved [Lieberam-Schmidt, 2010]. The aim of clustering is to group similar objects and separate them from less similar ones [Manning et al., 2008]. Clustering the results of browsing the RDF store was used to evaluate the ontology-based browsing that is based on the enrichments added to the RDF store by the LEMO system. The clustering techniques are unsupervised techniques for detecting groups of similar objects or a community of objects. In this section, a community refers to a community of EMOs that can be defined as a group of densely linked EMOs representing EMOs that have a common topic or subject [Liu, 2007]. As explained in this chapter, the ontology-based browsing method is based on the Term resources describing the subjects of the EMOs.

Validating the browsing results of clustering is implemented using the Agglomerative Hierarchical Clustering method to analyse the strength of the linkages between the collection of EMOs retrieved while browsing denoted as G. Since ontology-based browsing is based on the ontology classes and its hierarchies, hierarchical clustering methods are the best to define the hierarchies of nested clusters resulted in the browsing results [Ceri et al., 2013*a*]. In this method, clusters are created by merging the most similar points in the dataset into clusters based on a distance matrix and stops when all the clusters have merged into one big cluster. The results of merging the clusters are represented by a dendrogram, which represents the detailed merging process comprising a hierarchical tree of clusters. At some level in the tree, some meaningful clusters might be found. Several measures can be used to decide on the best number of clusters. The silhouette coefficient is used in this experiment since it assesses both the separation and cohesion of clusters [Granichin et al., 2015]. When clustering the datasets in this experiment, the average silhouette coefficient is calculated at different levels of clustering to determine

the best number of clusters for each experiment conducted. The detailed results of clustering experiments are explained after describing the data-processing needed to perform clustering and the distance function used. Also, the clustering analysis criteria applied in these experiments are explained in order to be used for validating and comparing the results of the clusters discovered at different branches while browsing. It is hard to evaluate the results of clustering large datasets. Therefore, clustering the browsing results is used for evaluating the separation and compactness of the groups of EMOs retrieved. The clusters visualized in Figure 7.9 indicate how the EMOs in the RDF store can compose one large complex connected graph. The figure illustrates the results of an experiment conducted for clustering a large sample of the EMOs in the RDF store. The EMOs comprising the LEMO RDF store number more than 10,000 EMOs. Hence, detecting communities based on the relations between these EMOs is not efficient. The figure illustrates some clusters

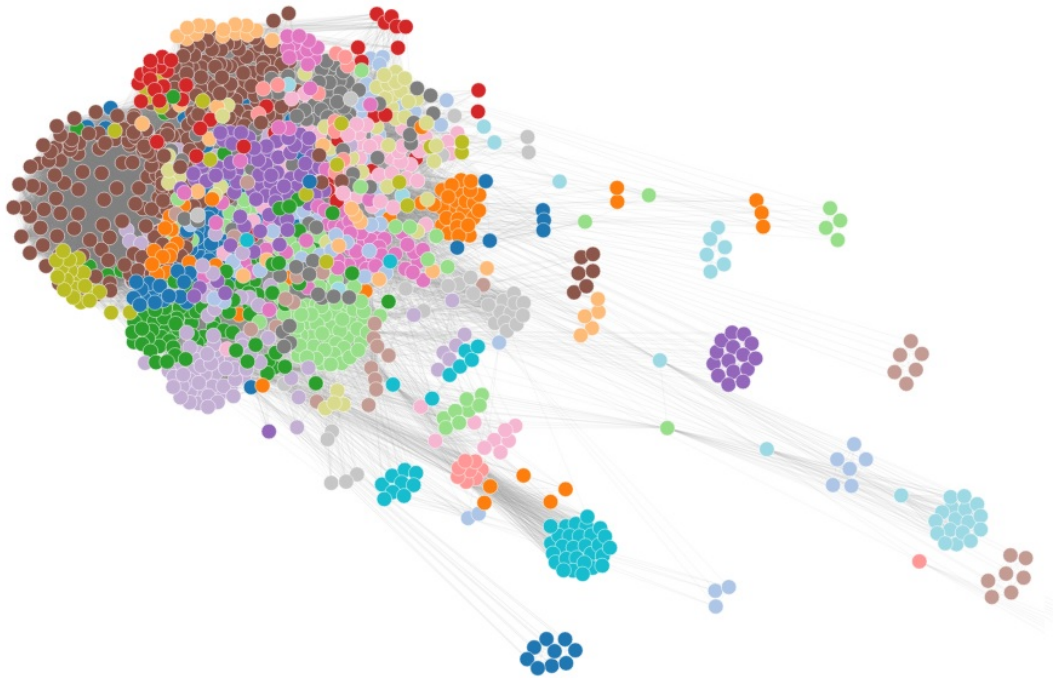


Figure 7.9: Sample of EMOs clustered from the LEMO RDF store

of EMOs but not all EMOs have been clustered in well-separated groups.

7.4.1 Data Pre-processing

In order to perform clustering on the browsing results referred to as G, the results of interlinked EMOs represented by the graph $G = (V, E)$ are converted into a matrix S. The matrix is considered a similarity matrix of EMOs where V is the set of EMOs retrieved and E are edges linking these EMOs. If the size of the vertices making the graph $|V| = n$, then the size of the similarity matrix built should be of a size $n \times n$. The values of the matrix represent the similarity between EMOs are the number of similar categories describing their Subject in the metadata. The matrix value is given as:

$$S = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \tag{7.2}$$

where x_{ij} represents the similarity between two EMOs based on the common Term resources related to its subject attribute. For example, the similarity between two EMOs emo_i and emo_j is equal to x_{ij} where $x_{ij} = |KS_i \cap KS_j|$.

7.4.2 Distance Function

After the data pre-processing of any retrieved result named G from browsing the RDF store, the results and their relations can be represented as a matrix to perform clustering. In the agglomerative hierarchical clustering method, clusters are created by merging the most similar EMOs into clusters based on a distance matrix that represents the distance between the EMOs. The distance matrix is built usng the values in the similarity matrix S (equation 7.2) for any results retrieved as $G = (V, E)$. where $V = \{x_1, x_2, \cdots, x_n\}$, edges $E = \{(x_i, x_j) \,|x_i, x_j \in V\}$. The distance

matrix W represents the pairwise distances among the n EMOs in the graph G. The matrix W is an $n \times n$ matrix given as:

$$\mathbf{W} = \left\{ \delta(x_i, x_j) \right\}_{i,j=1}^{n} \tag{7.3}$$

where

$$\delta(x_i, x_j) = \|x_i - x_j\| \tag{7.4}$$

is the Euclidean distance between x_i, x_j. That is, the distance weight $w_{ij} = \mathbf{W}(i, j)$ for all $x_i, x_j \in V$.

Based on the pairwise distances between EMOs, the clusters are created based on merging the most similar EMOs into one cluster, and the process iterates until one all the EMOs are merged into one big cluster. The steps of merging the EMOs are represented in a tree-like view called a dendrogram which represents the merging process comprising a hierarchical tree of clusters. In this experiment, the silhouette coefficient is used to decide on the best number of clusters. This measure assesses the separation and cohesion of clusters [Granichin et al., 2015]. Several internal measurements are explained in the next section to evaluate the results of the clustering experiments.

7.4.3 Clustering Analysis

Clustering the browsing results is performed to detect communities of related EMOs. Several measurements exist for validating the healthiness of the clusters discovered. In this research, we compare the healthiness of the clusters discovered in the results of browsing different branches of the LEMO navigational menu. The clusters or communities discovered in the results of browsing one branch in the navigational menu should be less separated than clusters discovered in the results of browsing different branches. In other words, the clusters discovered in one branch represent sub-topics of one wider topic, while the clusters of different branches represent distinct topics.

Hence, the quality of the clusters in the first case must be lower than the quality of the second case as they consist of well-separated EMOs, i.e., not related to each other. To perform the comparison between the results of the clustering experiments, some internal measurements for calculating the clusters quality are explained [Zaki and Wagner, 2014]. The internal measures utilise the notion of intracluster similarity contrasted with the notion of intercluster separation. Therefore, the next section explains the calculations important for understanding the internal measures.

Preliminaries

Given a clustering $C = \{C_1, C_2, \cdots, C_k\}$ where k is the number of clusters, and given any subsets $P, R \subset V$, define $W(P, R)$ as the sum of the weights w on all edges with one vertex in P and the other in R, given as

$$W(P, R) = \sum_{x_i \in P} \sum_{x_j \in R} w_{ij} \tag{7.5}$$

Also, given $P \subseteq V$, we denote by $\overline{P}$ the complementary set of vertices, that is, $\overline{P} = V - S$.

The internal measures are based on various functions over the intracluster and intercluster weights. In particular, note that the sum of all the intracluster weights over all clusters is given as

$$W_{in} = \frac{1}{2} \sum_{i=1}^{k} W(C_i, C_i) \tag{7.6}$$

It is divided by two because each edge within C_i is counted twice in the summation given by $W(C_i, C_j)$. Also, note that the sum of all intercluster weights is given as

$$W_{out} = \frac{1}{2} \sum_{i=1}^{k} W(C_i, \overline{C_i}) \tag{7.7}$$

Here too was divided by two because each edge is counted twice in the summation across clusters. The number of distinct intracluster edges, denoted N_{in}, and intercluster edges, denoted N_{out}, are given as

$$N_{in} = \frac{1}{2}\sum_{c=1}^{k}\sum_{x_i \in C_i}\sum_{x_j \in C_i} l(x_i, x_j), \quad l = \begin{cases} 1 & \text{if } e(x_i, x_j) \in E \\ 0 & \text{if } e(x_i, x_j) \notin E \end{cases} \tag{7.8}$$

$$N_{out} = \frac{1}{2}\sum_{c=1}^{k}\sum_{x_i \in C_i}\sum_{x_j \notin C_i} l(x_i, x_j), \quad l = \begin{cases} 1 & \text{if } e(x_i, x_j) \in E \\ 0 & \text{if } e(x_i, x_j) \notin E \end{cases} \tag{7.9}$$

Note that the total number of distinct pairs of points N is

$$N = N_{in} + N_{out} \tag{7.10}$$

Based on the values of the previous functions applied to the set of clusters C, multiple internal evaluation measures can be computed. To evaluate the quality of the clusters detected, the internal measures utilised the intracluster similarity and interclusters separation which are calculated in the functions explained before $W_{in}, W_{out}, N_{in}, N_{out}$.

Following is an explanation of the internal measures used for validating the clustering experiments, followed by the discussion of the clustering results and visualization of the clusters performed in this chapter.

BetaCV Measure

The BetaCV measures the quality of the clusters generated based on the ratio between the intracluster and intercluster distances. Equation 7.11 shows how the BetaCV is measured.

$$BetaCV = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} = \frac{N_{out}}{N_{in}} \frac{\sum_{i=1}^{k} W(C_i, C_i)}{\sum_{i=1}^{k} W(C_i, \overline{C_i})} \tag{7.11}$$

It evaluates the mean intracluster distances to the mean intercluster distances. The smaller the BetaCV ratio, the better the clustering. It indicates that, on average, the distances between points in the same cluster are smaller than distances between points in different clusters.

Normalized Cut (NC) Measure

The normalized cut measure can be used in the clustering process to determine the best cut for cluster partitioning. It can also be used as an internal measure of cluster quality. As for all the measures, we apply equation 7.12 on the distance matrix **W**. The value of NC is maximized when the intracluster distances are much smaller compared to the intercluster distances.

$$NC = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{W(C_i, V)} \tag{7.12}$$

where the volume of cluster C_i, denoted as $W(C_i, V) = W(C_i, C_i) + W(C_i, \overline{C_i})$. so that

$$NC = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{W(C_i, V)} = \sum_{i=1}^{k} \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \overline{C_i})} + 1} \tag{7.13}$$

The higher the normalized cut value, the better. The NC value is maximized when the ratio between the intracluster distances and the volume of the cluster are as small as possible across all the k clusters.

Modularity

The modularity objective for graph clustering is used as the third internal measure calculated using equation 7.14. The modularity measures the difference between the

actual and expected distances within the clusters.

$$Q = \sum_{i=1}^{k} \left(\frac{W(C_i, C_i)}{W(V, V)} - \left(\frac{W(C_i, V)}{W(V, V)} \right)^2 \right) \tag{7.14}$$

The equations are based on the distances matrix. Therefore, the smaller the modularity measure, the better the clustering. It indicates that the intracluster distances are low compared to the expected distances to the other clusters.

Davies-Bouldin (DB) Index

This measure is based on the cluster mean and variance values, and measures the quality of cluster separation [Zaki and Wagner, 2014]. Let μ_i denote the cluster C_i mean, given by

$$\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j \tag{7.15}$$

Further, let the variance σ_i denotes the spread of the points around the cluster mean defined in equation 7.16.

$$\sigma_i = \sqrt{\frac{\sum_{x_j \in C_i} \delta(x_j, \mu_i)^2}{n_i}} \; = \; \sqrt{var(C_i)} \tag{7.16}$$

Where $\delta(x_j, \mu_i)^2$ is the squared Euclidean distance for the points in the cluster from the cluster mean. Then, the Davies-Bouldin measure for pair of clusters C_i and C_j is defined in equation 7.17.

$$DB_{ij} = \frac{\sigma_i + \sigma_j}{\delta(\mu_i, \mu_j)} \tag{7.17}$$

DB_{ij} indicates how compact the clusters are compared to the distance between their means. Based on the DB_{ij} values for all pairs of clusters, the Davies-Bouldin Index

is defined in equation 7.18

$$DB = \frac{1}{K} \sum_{i=1}^{k} \underset{j \neq i}{max} \{DB_{ij}\} \tag{7.18}$$

The smaller the DB value the better the clustering. The index is calculated based on the largest DB_{ij} ratio for each cluster C_i. Hence, it will give a good indication about how well the clusters are separated from each other.

Following is a detailed explanation of the clustering results after conducting several experiments for clustering the results of browsing the nine dense nodes in the navigational menu at different levels of browsing. Visualization of the clustering results is necessary to demonstrate the strongly connected communities in the complex network of LEMO.

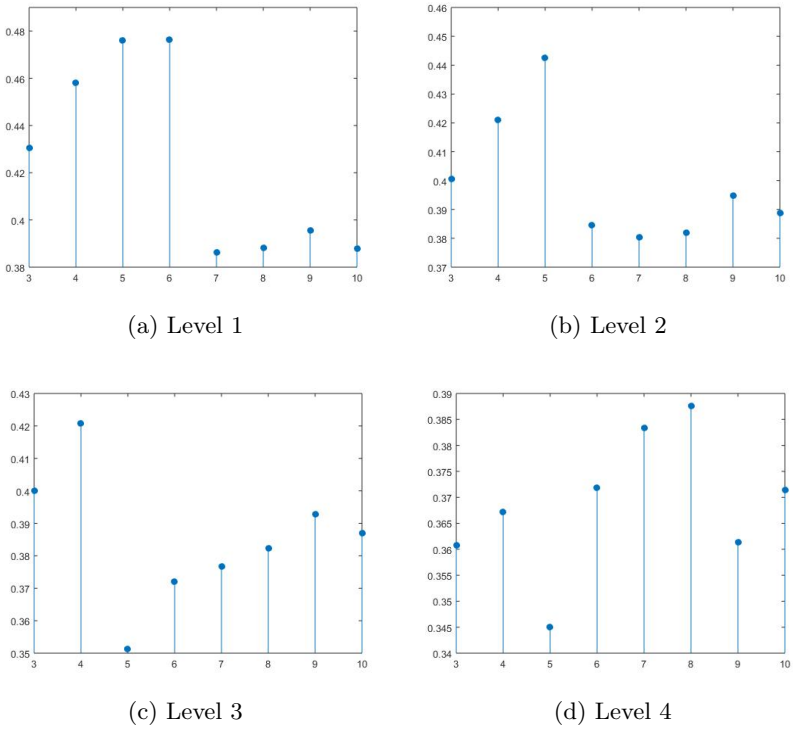
7.4.4 Clustering Experiments

Browsing the LEMO navigational menu in the previous section has indicated that browsing some concepts resulted in more dense results than other nodes. The results were detailed in Figure 7.6 and showed that nine nodes were considered for further evaluation. These nine node nodes where then browsed over several levels and the results were illustrated in Figure 7.8. It has been mentioned that the results of browsing these 9 nodes can be used for clustering experiments were the dataset retrieved are dense and can produce efficient clustering. The full results of clustering experiments of these nine nodes are detailed in Table 7.1. The table details the results of applying agglomerative hierarchical clustering on the retrieved EMOs datasets of size (m) at each trial. At different levels, the highest silhouette value (s) and its associated number of clusters (k) are listed. The average silhouette value for each trial and the maximum links density value for the EMOs retrieved at different levels for each node.

As explained in the previous sections, the silhouette coefficient is a measure

Table 7.1: Experiment of clustering results of browsing the branches at different levels

levels / nodes	**level 1**			**level 2**			**level 3**			**level 4**			**level 5**			Avg(s)	Link Density
	m	s	k	m	s	k	m	s	k	m	s	k	m	s	k		
node 1	6950	0.475	4	6134	0.468	5	2784	0.462	4	801	0.427	3	568	0.462	3	0.459	0.464
node 2	6732	0.478	6	4932	0.442	5	4776	0.420	4	3881	0.385	8	2493	0.408	9	0.427	0.424
node 3	8495	0.529	5	7055	0.552	8	4820	0.548	3	4585	0.554	3	1297	0.549	3	0.546	0.556
node 4	7977	0.477	5	6919	0.466	5	6907	0.465	5	5405	0.443	5	1760	0.361	9	0.442	0.527
node 5	3748	0.5	3	2551	0.508	3	2051	0.509	3	2051	0.509	3	1934	0.506	3	0.506	0.449
node 6	2512	0.44	4	513	0.342	7	245	0.37	5	224	0.4	3	215	0.385	3	0.387	0.459
node 7	1049	0.499	8	341	0.449	6	24	0.387	6	24	0.387	6	-	-	-	0.431	0.394
node 8	2305	0.536	5	1541	0.481	5	1368	0.47	5	990	0.457	5	457	0.518	6	0.492	0.502
node 9	135	0.342	7	59	0.41	6	59	0.41	6	59	0.41	6	-	-	-	0.393	0.452



(a) Level 1 (b) Level 2

(c) Level 3 (d) Level 4

Figure 7.10: The silhouette plots for node 2 (Substance branch) while clustering

of both the cohesion and the separation of clusters. The agglomerative hierarchical clustering method clusters the data in iterations, and at every iteration, the number of the clusters changes due to merging the most similar EMOs at each iteration. Calculating the silhouette value for different iterations helps decide on what is the best number of clusters that results in more compact and separated clusters. The number of clusters k is the best number of clusters based on the highest silhouette value calculated. For example, for the clustering experiments conducted on node 2, that represent the ontology concept (Substance), detailed in Table 7.1, the results of the silhouette values associated with the clustering experiments conducted at different levels are illustrated in Figure 7.10. The number of clusters k with the highest silhouette value is chosen as the best number of clusters for the dataset tested.

Now that the silhouette coefficient is explained, the results of the clustering

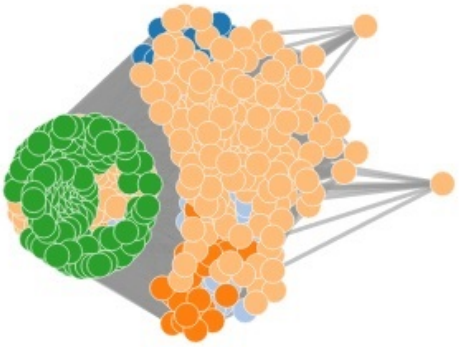
experiment detailed in Table 7.1 are discussed. The clustering is conducted on the results of browsing the densest branch in each node. The nodes are detailed in Figure 7.8 and have been named from node 1 to node 9 in the same order as detailed in the graph for easier reference. The simulations of browsing the RDF store has mimicked the user clicks on every ontology concept presented in the LEMO navigational menu. From the results of this simulation, clustering experiments were conducted only on dense branches of these nine nodes. In other words, the node that retrieves the highest number of EMOs in level 1 is chosen for each of these nine nodes, and then for level 2 of each node, the descendent node with the highest number of EMOs retrieved is selected for browsing and clustering its results, and so on. As a result, the simulation actions can be visualised as browsing one branch of every node. Table 7.1 details the results of the clustering experiments conducted on each branch of the nine nodes.

It is noticed that clustering is more efficient with larger datasets retrieved compared to smaller ones. The largest dataset was retrieved when browsing node 3 (Clinical finding) and resulted in more than 8000 EMOs retrieved which are annotated with descendent classes. The clustering consistently works well at deeper levels too. Node 3 has the highest average silhouette value of all the clustering performed at different levels while the lowest average silhouette value is related to node 9 (Specimen) which has the lowest number of EMOs retrieved. The clustering experiment gives a good indication for the distribution of the LEMO dataset over the SNOMED CT ontology resulting from the enrichment process. At some branches, the nodes did not extend to more than four levels as in the cases of nodes 7 and node 9. This experiment does not give any validation of the correctness of the linkages made in LEMO based on subject annotations. It only indicates that at different levels of browsing, although the data retrieved in deep levels for a subset of those in the higher levels, the results of the clustering will change and in some cases improve. For further discussion, the node with the highest average silhouette

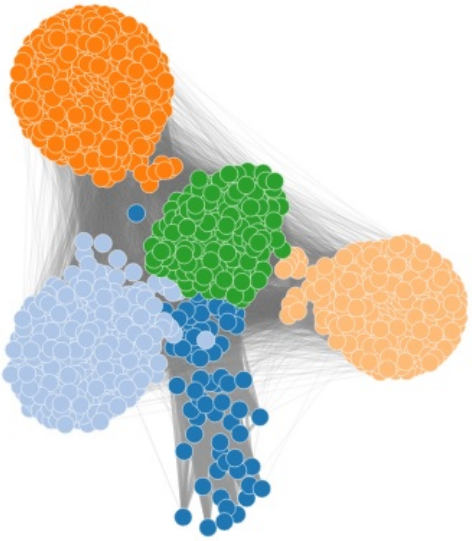
value selected as a case study is node 3. The clusters discovered in this node should be of a better quality than other clusters. Therefore, it was selected to perform the following comparison.

A comparison between two sets of clusters was conducted to validate the efficiency of the ontology-based browsing method used for retrieving EMOs from the RDF store. The comparison is performed on clusters detected in the results of browsing one node that is the densest node in the ontology (node 3). Clustering the results of browsing this node at a deeper level that is level 6 had 5 clusters detected. These five clusters are considered the first set in the comparison and named the **node clusters**. The second set of clusters consists of 5 clusters detected while clustering the results of browsing deep levels of different branches named, and this set is named the **branches clusters**. The two sets of clusters are chosen to be of the same size to conduct a comparison between their internal measures calculated. The results of browsing node 3 at level 6 resulted in having 568 EMOs retrieved. After performing Agglomerative Hierarchical clustering on these EMOs, the results were 5 detected clusters illustrated in Figure 7.11. The figure illustrates the 5 clusters in different colours and their associations in order to evaluate the separation and compactness of the clusters. Browsing in deeper levels in the navigational menu results in smaller number of EMOs retrieved and that is more efficient for visualizing the linkages between the EMOs.

The **node clusters** are not well separated as detailed in Figure 7.11a. The links between EMOs are strong making the visualization of different clusters hard. Although the data are strongly linked, the clusters are compact but not well separated. This visualization for the dataset content is logical since all the EMOs are retrieved from one branch at a deep level that is level 6. The visualization of the EMOs and their relations in the second set, **branches clusters**, is illustrated in Figure 7.11b. It consists of 1322 EMOs separated into 5 clusters detected when clustering the results of browsing different branches. The clusters were selected ran-



(a) The node clusters



(b) The branches clusters

Figure 7.11: Visualization of the two sets of clusters

domly from the sets of clusters detected in the experiments of clustering the results of browsing nodes at level 5. The clusters illustrated in Figure 7.11b are well separated from each other and each cluster is compact and has weaker linkages to other clusters.

To support the visualization of the clusters composing the two sets in Figure 7.11, a comparison between the internal measures indicating their quality is explained. The results of calculating the four measurements are: BetaCV measure,

Table 7.2: Comparison of the evaluation measures

	internal evaluation measures			
Datasets	BetaCV	NC	Modularity	DB
Node dataset	0.6115	0.8567	-0.0144	0.1116
Branches dataset	0.4339	0.8938	-0.0216	0.0621

Normalized Cut (NC) measure, the Modularity, and the Davies-Bouldin (DB) measure are detailed in Table 7.2. The results indicate that the clusters in the **branches clusters** are more compact and separated than the clusters in the **node clusters**. The lower the values of the BetaCV, Modularity, and Davies-Bouldin index, the better the clusters are in its separation and compactness, while higher values for NC indicate better clusters as explained in the previous section. The results are fairly close to each other, suggesting that the two sets of clusters are well-separated from each other and compact at the same time. The results indicate the ontology-based browsing was able to retrieve EMOs that are related to each other and have similar topics based on their Term resource annotations. The comparison results support this conclusion since the clusters of different branches are considered of a better quality than clusters of one branch as detailed in the table. Further experiments can be conducted with expert users in the future to evaluate the linkages discovered between samples of EMOs collected from the LEMO RDF store.

7.5 Summary

Information retrieval techniques are crucial for any information system that are browsing and query searching techniques. As for the LEMO dataset, browsing its content plays a vital role in discovering the topics covered in this dataset and the quality of the linkages generated between the EMOs composing it. In order to explore the LEMO dataset, the features introduced in the LEMO metadata schema for describing annotations discovered in the EMOs metadata has been exploited for designing the browsing methods. This chapter has presented techniques developed

for building navigational menu for accessing the LEMO dataset stored in an RDF store. Furthermore, it has examined the results of browsing the RDF store based on the ontology concepts annotating its EMOs.

The proposed browsing model is based on the hierarchical relations in the SNOMED CT ontology that is used to enrich the LEMO dataset. In particular, this chapter evaluated the use of SNOMED CT concepts for indexing and categorising the components of the LEMO dataset. In order to measure the effectiveness of the proposed browsing model, this chapter has conducted experiments for clustering the browsing results based on the similarity of the EMOs retrieved. The EMOs are considered as vectors of ontology concepts that are used to annotate its content, and the clustering experiments grouped the related EMOs based on the vectors distances. The results presented in this chapter have indicted that ontology-based browsing has succeeded in indexing the similar EMOs using ontology concepts used to enrich its description.

In conclusion, this chapter has addressed the research objective **O9**: "Develop ontology-based method for browsing the LEMO dataset resulted from the LEMO system and evaluate the similarity between the results retrieved while browsing". The process of addressing this research objective has partially contributed in answering research question **R4**: "How can the Linked Data practices be utilised in the process of accessing and querying the dataset of integrated EMOs called the LEMO dataset? Moreover, how can the linkages between content retrieved from the LEMO dataset be evaluated?". The answer for the utilising the Linked Data for accessing the LEMO dataset has been explained in the results of the experiments conducted to test the proposed ontology-based browsing model. To complete the answer of this research question, the following chapter explains how Linked Data can be utilised to build an ontology-based query searching method that enhances the traditional text-based search in the LEMO dataset.

Chapter 8

Information Retrieval by Query Searching

8.1 Introduction

So far in this research, methods and techniques have been proposed for building a linked dataset of Educational Medical Objects (EMOs) from diverse web data sources named the LEMO dataset. An RDF store is used to store and manage the content of the LEMO dataset (chapter 6). In order to access and retrieve EMOs from the RDF store, specialised methods were designed for browsing and querying this RDF store. The ontology-based browsing method has been developed and evaluated in the previous chapter (chapter 7). Query searching is considered an important information retrieval technique which is usually associated with any IR system [Zhang, 2008*b*]. It is an information retrieval method used to satisfy the user's information needs. The users are seeking to answer their needs. Hence, they formulate a query in the specific query language and the system then returns the documents which match the query. Robust information retrieval systems, such as existing search engines, have been developing and applying a well-known algorithm for processing user queries over large information spaces [Lieberam-Schmidt, 2010].

Moreover, information retrieval models such as the Boolean model [Manning et al., 2008] and vector space model [Liu, 2011] have proved their efficiency in retrieving results that are relevant to a user query. Hence, applying such models for querying the RDF store guarantees effective retrieval of results as any other IR system. However, the LEMO metadata schema proposed in this research introduces new elements for describing the EMOs that are stored in the RDF store, which can be utilised for query searching. Ontology concepts were used to annotate the EMOs descriptions represented as Term resources in the RDF store. This research aims at exploiting the ontology concepts used to describe the EMOs for designing the query searching method. Based on existing information retrieval models, an ontology-based query searching algorithm is proposed and implemented to search the RDF store. A web access interface is designed to perform the query searching on the RDF store. As the RDF store is not published for public access yet, the query searching interface was implemented on a local server to test and validate the algorithm proposed. By simulating user behaviour, several experiments were conducted for searching the RDF store. The results of applying this algorithm for when searching was compared with traditional text-based searches in the RDF store.

The web access interface was developed with SPARQL query language. It is the W3C Recommendation for an RDF query language that supports querying of multiple RDF models [Quilitz and Leser, 2008]. The RDF store represents EMO metadata as a collection of related RDF resources, and the ontology-based searching algorithm proposed, suggests utilising the RDF resources and their relations to enhance the discovery of EMOs. The process of developing and testing the ontology-based search algorithm proposed is explained in this chapter. Since the LEMO dataset is not published for public access at this point in the research, one way to evaluate the proposed query searching method is by running random queries performed by a system. Simulation of user behaviour is beneficial for replacing possible queries that might be initiated by real web users with simulated queries

performed by the system. Query searching evaluation can be simulated as it can be done with considerably less time and efforts compared to user testing. Also, the RDF store is built with no human involvement when harvesting its content, therefore, having EMOs which satisfy the needs of all user queries is hard to achieve. Thus, the evaluation has experimented with a variety of assumptions about possible user queries, controlled by the content of the RDF store.

8.1.1 Chapter Objectives

For the sake of evaluating the discovery of EMOs in the RDF store, this chapter aims at presenting an ontology-based query searching method, its implementation, and the needed experiments for evaluating its results. It addresses the research objective **O10**: "Develop an ontology-based query searching algorithm for testing and comparing of query searching results between ontology-based and text-based searching methods in the LEMO dataset". Possible users' actions for query searching were simulated and applied to demonstrate the results of the ontology-based query method over the RDF store. The query searching results were compared with the results of simple text-based query searching. The goal is to have a query searching method that enhances the discovery of EMOs and expands the search results based on the ontology concepts categorising these EMOs. In what follows, a scenario describing the simulated process of sending a query, processing it, and retrieving EMOs matching that query. The scenario describes the proposed method of query searching based on the SNOMED CT ontology.

8.1.2 Scenario Example

Simulating the query searching process is the evaluation technique followed in this chapter. Hence, an example is given in this scenario for possible user behaviour query searching the RDF store.

When users is seeking information about *"Kidney Disease"*, they are prompted

with an auto-complete search box. The text searched is bound to the Classes resources that represent the ontology classes as stored in the RDF store. The proposed ontology-based query searching method consists of two steps. First, it starts with enriching the query initiated using related ontology classes which describe the context of the concept queried. The second step is matching the query vector with EMOs stored in the RDF store. For this example, the searched word is bound to the Class resource that represents the SNOMED CT concept *"Kidney Disease"*[1]. Behind the scene, a query vector is built exploiting the predicates describing this Class resource in the RDF store. The Class resource describes its adjacency with related classes as it is represented in the SNOMED CT ontology hierarchy. As a result, a query vector is built from a collection of classes related to the queried ontology concept. Further processing of the collection of classes is performed to assign weights to the vector classes that indicate the strength of the relation according to its co-occurrence in the RDF store. The query vector is then matched with the Term resources annotating the EMOs. Each EMO is represented as a vector of the Term resources related to it, and the weights of the Term resources are used for calculating their relevance to the query vector. The results of the search method are retrieved and ranked after calculating the distance between each EMO vector and the query vector that has resulted from expanding the ontology concept queried. The detailed process of the ontology-based query searching is explained in the rest of this chapter.

8.1.3 Chapter Outline

This chapter is outlined as following. Firstly, the development of the web access interface for performing query searching is detailed. Then, the steps involved in the proposed ontology-based query searching method are described that include query expansion process and the matching and ranking process. Finally, the experiments

[1] http://purl.bioontology.org/ontology/SNOMEDCT/90708001

conducted for testing and evaluating this proposed query searching method are detailed. Some evaluation criteria are explained to use them for proving the efficiency of the proposed method compared to simple text matching query searching. The experiments aim to highlight the enhancement in discovering more EMOs compared to traditional search methods such as simple text matching.

8.2 Ontology-based Query Interface

The web access interface developed in the LEMO system is illustrated in Figure 8.1. It is developed as a prototype in this research and will be incorporated in full LEMO website in the future. The search input box allows the user to enter an ontology concept by providing an auto-complete feature that binds the user with a collection of ontology concepts stored in the RDF store. As shown in the text search box, the user is prompted with suggestions based on their input in the text box. Since the query searching method proposed is based on ontology concepts, the collection of Class resources stored in the RDF store binds this auto-complete search box based on matching the classes labels with the user entries. The Class resources represent the Term resources used to annotate the EMO resources with SNOMED CT ontology concepts.

To formally represent the process behind this interface, the design of the RDF store must be considered (Chapter 6). The collection of Class resources and their relations can be represented in a graph structure denoted by the graph G_{emo}. The auto-complete search box binds the user with the Class resources by matching the text they input with the Class label stored in the description of the Class resource description. Hence, the auto-complete search field retrieves the values from the `<lemo:lemoTermClassLabel>` property of the Class resource (chapter 6, section 6.2). The following SPARQL query, shown in Listing 8.1, is initiated by the auto-complete text box, and all the ontology classes' labels matching this SPARQL query are

retrieved from the RDF store. Clicking on one of these ontology concepts' labels retrieves the URI of that Class resource that will be considered as the value of the search box and used retrieving content from the RDF store. The user input in the auto-complete search box is passed to this SPARQL query via the variable $\$term$. The auto-complete search is case insensitive, and the results are ordered alphabetically based on the ontology classes labels as detailed in the query.

Listing 8.1: Auto-complete query

```
PREFIX lemo:<http://www.warwick.ac.uk/ias/lemo/> .
             SELECT ?id ?label WHERE {
             ?id lemo:lemoTermClassLabel ?label .
             Filter(REGEX(?label, '^$term', 'i'))
             }
             Order by ?label
```

After selecting an ontology concept using the auto-complete search box, the user can click the search button to search the RDF store and retrieve EMOs that match their query. Clicking the search button retrieves a collection of relevant EMOs

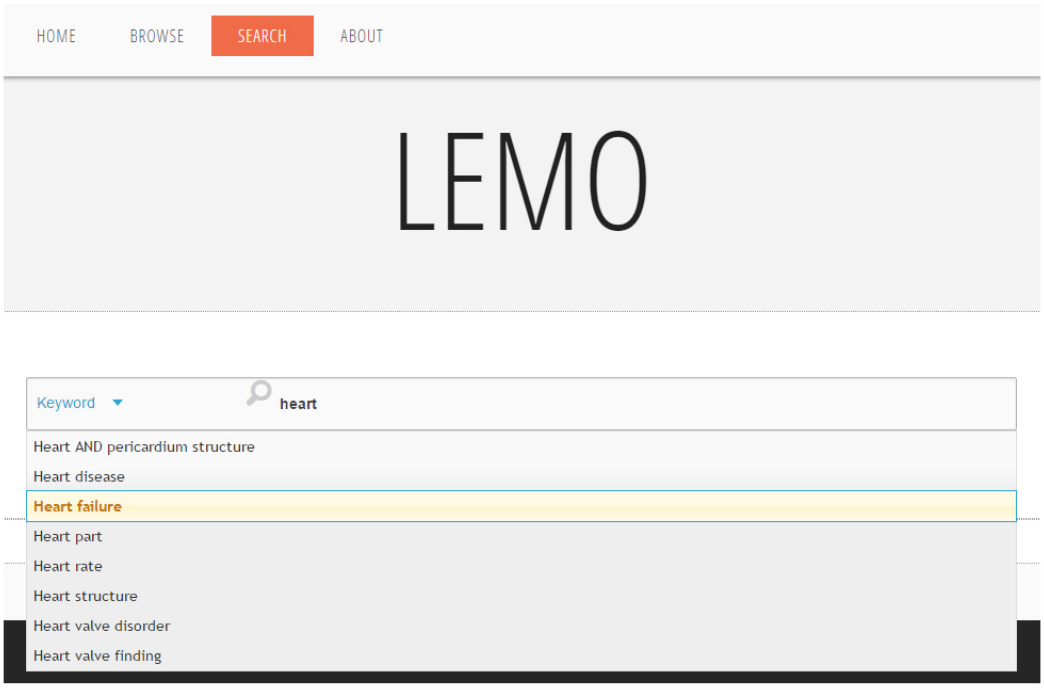


Figure 8.1: The ontology-based LEMO search user interface

from the RDF store and display them as a list. A sample of the search results for searching the concept *"Heart failure"* is represented in Figure 8.2 after selecting the concept using the query interface illustrated in Figure 8.1 . The list of search results shown in the figure details the title, description, and type of four EMOs

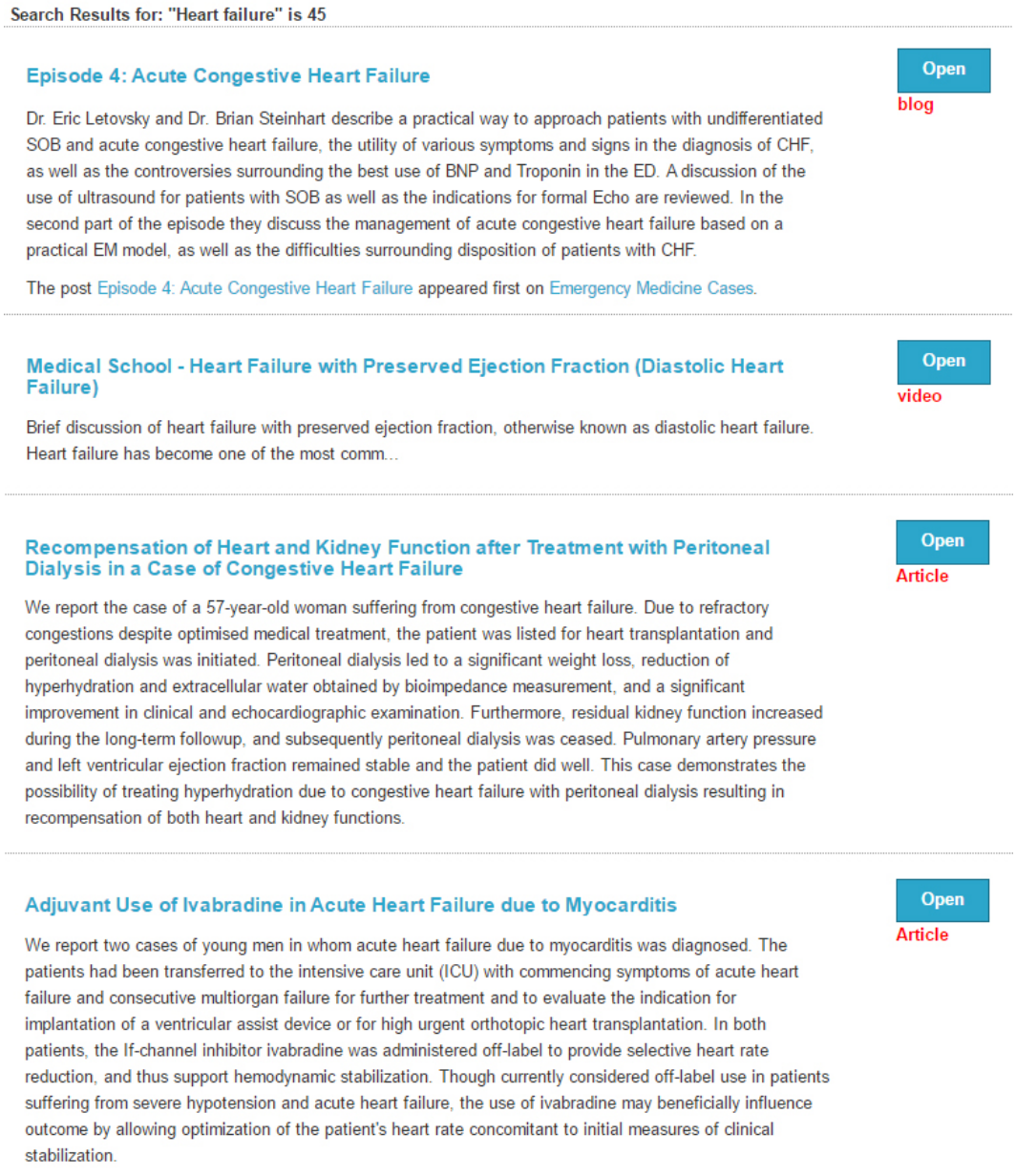


Figure 8.2: Query results for "Heart failure"

retrieved from the RDF store. The total number of EMOs retrieved is 45 EMOs as indicated in Figure 8.2. The detailed process of matching and retrieving these search results are explained in the following section. The open button shown next to each EMO navigates the user to the original website from where the EMO was harvested. The RDF store is considered as a metadata portal that stores all the EMOs metadata and their URLs. The URLs linking each EMO with its original website are stored in the `<dc:identifier>` metadata element. Further metadata elements can be retrieved from the RDF store to include in the results description such as creators, terms annotated, and the publisher.

8.3 Ontology-based Query Searching

A major challenge in implementing query searching to search the RDF store is retrieving relevant EMOs, which are of different types and harvested from various web stores. Overcoming this challenge is particularly the aim of developing the LEMO system. The proposed method of ontology-based query searching consists of two steps. It starts with query expansion to extend the ontology concept queried into a vector that represents the context of this ontology concept. The second step is matching the query vector for results from the RDF store. The query searching method is proposed in algorithm 4 and the two steps are explained in details in the following sections. The algorithm explains the steps of performing a query search on a local server as the RDF store is not published for user access. The two processes involved in the query searching are performed in SPARQL queries as the search interface is implemented as a web interface.

8.3.1 Ontology-based Query Expansion

After selecting an ontology concept using the auto-complete search text box provided by the query interface, the ontology class representing that concept is ex-

Algorithm 4 Ontology-based Query

Input : Ontology class to be queried q_0,
Output : Ranked Search Result set of EMOs R

procedure BUILDQUERYVECTOR(q_0)
  $Q \leftarrow getRelatedClasses(q_0)$ ▷ Stores adjacent classes to q_0
  for $c \in Q$ **do**
    $qResults \leftarrow getEMOsAnnotatedWith(c)$

    add $qResults$ to $Rtemp$ ▷ Rtemp is the final search results
  end for
  $QW \leftarrow$ weightQVector(Q) ▷ Weight related classes to q_0
end procedure

procedure MATCHANDRANK(QW)
  for $emo_i \in Rtemp$ **do**
    $T_i \leftarrow weightDVector(emo_i)$ ▷ Weight emo_i annotations based on QW
  end for
  for $emo_i \in Rtemp$ **do**
    calculatedEuclideanDist(T_i, QW)
  end for
  R $\leftarrow$ Sort($Rtemp$) ▷ Sort results ascendingly
end procedure

panded into a query vector. The query vector is built based on the ontology class hierarchy defined in the SNOMED CT ontology and described in the RDF store. Each ontology class is represented as Class resource in the RDF store and described using predicates for defining adjacency relations, `<lemo:adjacentTo>`, between the ontology classes. The collection of the Class resources in the LEMO dataset are represented as graph G_{emo} and the adjacency relations between these Classes are used to expand the queried ontology class into a query vector. The Classes that are directly connected to the query class forms the query vector that is used in the search process. These classes can either be a descendant class or an ancestor class directly related to the query class.

The process of searching for the ontology class *"Kidney disease"*, presented in the scenario example at the beginning of this chapter, starts with expanding the search process to a vector of classes related to that ontology class. Based on

the relations between the Class resources stored in the RDF store, a vector can be formed from the set of its related classes retrieved from reading the RDF statement describing the ontology class with the predicate `<lemo:adjacentTo>`. All the Class resources related to the ontology class *"Kidney disease"* compose the query vector. However, these Classes must be weighted to indicate the strength of their relation to the queried ontology class, *"Kidney disease"*, according to its co-occurrence in the RDF store. Some ontology classes might not be used to annotate any EMO in the LEMO dataset, and they are used as transition nodes to build the hierarchical relations of the ontology concepts. In this example, the related classes to the ontology class *"Kidney disease"* are 35 concept which are either parents or children of the query concept. Only 20 Class resources annotated EMOs in the RDF store, and the rest are created as transition nodes to build the relations between the ontology concepts to build the graph G_{emo}. After ranking and weighting the related classes against the EMOs in the RDF store, only 20 ontology classes were retrieved from the RDF store resources describing the annotations of EMOs. These classes have comprised the query vector in addition to the main ontology class *"Kidney disease"*. The final query vector and the weights assigned to its classes are represented in the Table 8.1.

The query concept that initiates any search query process is denoted as q_0. The collection of ontology classes related to the query class q_0 forms the vector Q given as:

$$Q = \{q_0, q_1, q_2, q_3, \cdots, q_j\} \tag{8.1}$$

In the example detailed above, the ontology class q_0 represents the *"Kidney disease"* ontology concept. The size of the query vector Q is denoted by M where $M = 21$ in this example. The weights assigned to Q are denoted by the vector QW of size M. It stores the weights of all the query classes detailed in Table 8.1.

The weights of this query vector are calculated based on the collection of EMOs R matching the query vector Q. The collection of EMOs forming R must have one of the ontology classes in Q annotated and represented Term resources related to in the RDF store, that is:

$$emo_i \in R, \quad if \quad |K_i \cap Q| >= 1 \tag{8.2}$$

The size of the collection of EMOs retrieved R is denoted by N, that is $|R| = N$. The weights of the query vector Q are calculated based on the percentage of its appearance in the query result R. For a query class $q_j \in Q$, the weight of the

Table 8.1: Classes in the query vector

Number	**Concept**	**Weight**
1	Kidney disease	1.000
2	Perinephritis	0.997
3	Hemorrhage of kidney	0.997
4	Renal tubular disorder	0.997
5	Renal abscess	0.994
6	Amyloid nephropathy	0.994
7	Hypertensive renal disease	0.991
8	Renal tubular acidosis	0.991
9	Infectious disorder of kidney	0.988
10	Obstructive nephropathy	0.988
11	Neoplasm of kidney	0.986
12	Renal cortical necrosis	0.986
13	Simple renal cyst	0.977
14	Papillary necrosis	0.974
15	Kidney stone	0.968
16	Glomerular disease	0.965
17	Renal impairment	0.954
18	Uremia	0.948
19	Nephrotic syndrome	0.936
20	Pyelonephritis	0.936
21	Injury of kidney	0.873

class $Weight(q_j)$ is given as:

$$Weight(q_j) = f_{q_j,R}/N \tag{8.3}$$

Where $f_{q_j,R}$ denotes the occurrences of the ontology class q_i in the Term resources annotating the EMOs in the search result R. If $R = \{emo_1, emo_2, emo_3, \cdots, emo_i\}$ then:

$$f_{q_j,emo_i} = \begin{cases} 1 & \text{if } q_j \in K_i \\ 0 & \text{if } q_j \notin K_i \end{cases} \tag{8.4}$$

Hence, $f_{q_j,R}$ is calculated by:

$$f_{q_j,R} = \sum_{i=1}^{i=N} f_{q_j,emo_i} \tag{8.5}$$

The weights for all the ontology classes in Q are calculated and stored in QW for matching the results and ranking them. The main ontology class q_0 is given a weight equal to 1 to emphasise its importance when ranking the results. The weights assigned to the rest of the query classes denote its association with the main ontology class.

8.3.2 Matching the Results

The query searching process was initiated by an ontology concept that represents a Class resource denoted as q_0. The ontology-based query searching method proposed expands this query class into a query vector Q that is weighted and stored in another vector QW both of size M. According to the proposed methodology, the EMOs retrieved in R can be ranked based on matching the query vector Q with the collection of Class resources annotating these EMOs. Therefore, each EMO $emo_i \in R$ is represented as a vector T_i of the same length of QW. The vector T_i represents the

weights of the ontology classes in Q if they exist in the set K_i related to emo_i. The weights of ontology classes in an EMO are based on the weight value assigned to Term resources related to it using the predicate `<rdf:value>`. If we can represent the EMOs as vectors with respect to the query vector, any distance measure may be used to rank the EMOs according to their relevance to the query. In representing an EMO as a vector, a weight must be assigned to each Term resource in the EMO that matches a query class in the query vector. The weight must represent the value of the frequency of that Term in respect to the collection of Term resources annotating that EMO. Many formulae for assigning these weight have been proposed in the literature. However, in vector-space modelling, these formulae are usually characterized by TF-IDF weights [Büttcher et al., 2010]. In the ontology-based query searching, each EMO emo_i consists of a collection of Term resources K_i and the goal of weighting the Term resources is to calculate the importance of one Term to the collection of Terms related to an EMO. Hence, it is possible to use the TF-IDF technique to determine the weights for Term resources in the vector representing the EMO before matching it with the query vector QW.

Therefore, each $emo_i \in R$ is represented as a weighted vector T_i where the size of $|T_i| = |Q|$ since the EMOs are weighted with respect to the query vector. The weights assigned to the vector T_i represent the weights of query classes in Q if they exist in the list of Terms annotated in emo_i. The weights of query classes Q are weighted for each $emo_i \in R$ and the weights are assigned to a vector T_i, such that:

$$weight(q_j, emo_i) = \begin{cases} TF(q_j, emo_i) \times IDF(q_j, emo_i) & \text{if } q_j \in K_i \\ 0 & \text{if } q_j \notin K_i \end{cases} \tag{8.6}$$

The $IDF(q_j, emo_i)$ function relates the term frequency of a query class an-

notated in an EMO to the total number of terms annotating that EMO, and the $TF(q_j, emo_i)$ indicate the term frequency of that query class if annotated in emo_i. There are variants of both functions have been proposed in the literature, but the following functions have been used in this research based on [Büttcher et al., 2010].

$$TF(q_j, emo_i) = \log(f_{q_j,emo_i}) + 1 \tag{8.7}$$

and

$$IDF(q_j, emo_i) = \log(\frac{|K_i|}{f_{q_j,emo_i}}) \tag{8.8}$$

Given that f_{q_j,emo_i} equals the value of the predicate `<rdf:value>` describing the Term resource that is annotated by the ontology class q_j. This value is calculated as part of enriching the RDF store explained while developing the LEMO system (chapter 5). The values of the weights must be normalized between 0 and 1 in order to produce valid results when compared with the query vector. Computing all the TF-IDF values for every $emo_i \in R$, the final results are represented as a set of vectors $T = \{T_1, T_2, T_3, \cdots, T_N\}$ that is $|T| = |R|$.

Now that both the query vector and the EMOs in the search results are represented by vectors QW and T, the set of search results R can be ranked based on the euclidean distance between QW and each $T_i \in T$ given by:

$$Euclidean(QW, T_i) = \sqrt{(QW - T_i)^2} \tag{8.9}$$

Based on the distance values of all EMOs in the search result retrieved, the EMOs can be ranked from the most relevant to the less relevant. The ontology classes forming the query vector can be represented as a small graph that is centred on the main ontology class q_0. Having an EMO retrieved that based on its distance from a query vector indicate that this EMO might be annotated with most of these

classes. Low distance values assure that the EMO is very related to the topic selected for querying.

To test the validity of this proposed method for query searching the RDF store, several experiments simulating user actions of searching were conducted. The simulations were performed by selecting random queries from the Class resources stored in the RDF store and used for annotating its EMOs. Before illustrating the results of the testing experiments, the following section presents the evaluation criteria used to validate the results and compares it with simple text matching search technique.

8.4 Evaluation

The proposed ontology-based query searching method has been tested by conducting several experiments to measure the efficiency of the method. The method aims to enhance the discoverability of EMOs stored in RDF store. The discoverability is improved in this ontology-based query searching by expanding the search query to include several concepts that represent the context of the query. However, the main query search that represents one ontology concept is emphasized by giving it the highest weight of all the other concepts resulting from expanding the search query. When the results are retrieved and ranked, although some of the EMOs retrieved might not include the main ontology class annotated, such results will end up ranked at the bottom of the list based on their Euclidean distance from the query vector.

The focus of this evaluation is to assess if the proposed method enhances the discoverability of the EMOs. It is compared with a simple text-based matching search. In traditional IR retrieval systems, the simplest method of searching is to use the Boolean model, also called the exact matching model. The query is presented in a boolean expression, and an EMO is retrieved if it satisfies the Boolean query expression. In these experiments, the goal is to test how efficient is

the proposed ontology-based query searching method compared to text matching techniques. Therefore, simple forms of queries are initiated to experiment with the ontology-based query searching by searching one ontology concept at a time. Searching more than one ontology concept requires the use of boolean expressions (using Boolean operators AND, OR, and NOT) that is not supported by the proposed method at this stage of the research. In the text-based query searching method, the ontology concept is searched by matching the label of the ontology concept with the textual description of EMOs, and the results of the two search methods are compared. At this point of the research, the simplest form of queries (that is one ontology concept, represented as a piece of text) is beneficial for testing and comparing the results of the two techniques. Expanding the functionality of the developed web search interface to include Boolean expressions for querying one or more ontology concept will be implemented in the future.

8.4.1 Preliminaries

The comparison between the query searching results of both methods, ontology-based and text-based search, is based on several evaluation criteria. To explain these evaluation criteria we need to define the sets of search results to compare. As detailed in the previous chapter, the search results of R represent the EMOs retrieved from the ontology-based method where $R = \{emo_1, emo_2, emo_3, \cdots, emo_i\}$. Assuming that the set of search result B is the set of EMOs retrieved from the text-based search, that is $B = \{b_1, b_2, b_3, \cdots, b_j\}$. Then, the total number of EMOs retrieved in the two methods is $|U| = n$ where $U = R \cup B$ and $U = \{u_1, u_2, u_3, \cdots, u_n\}$. The two sets must be represented as vectors of size n in order to calculate the similarity between the two search results, such that:

$$R_v = \{r_1, r_2, r_3, \cdots, r_n\},\ r_i \in \{0, 1\} \tag{8.10}$$

$$B_v = \{b_1, b_2, b_3, \cdots, b_n\},\ b_i \in \{0, 1\} \tag{8.11}$$

Where $r_i = 1$ if $u_i \in R$, otherwise $r_i = 0$. The same values applies to the vector B_v. For example, if $R = a, b, c$ and $B = d, f$ then, $U = \{a, b, c, d, f\}$, and the Table 8.2 represent the two vectors representing the results. The two vectors can be compared for similarity using different criteria. The table represents how to the two vectors R_v and B_v are used to represent the set R and the set B respectively. The evaluation criteria applied are discussed next.

Table 8.2: Example of two search results vectors

vectors	a	b	c	d	f
R_v	1	1	1	0	0
B_v	0	0	0	1	1

8.4.2 Similarity Measures

The process of retrieving the search results of the ontology-based query searching method proposed and the simple text-based query searching method have been explained. The following measurements are used to compare the search results stored in R and B for ontology-based and text-based methods respectively. Moreover, the vectors representing the two search results, R_v and B_v, are used to calculate the similarity. The vectors R_v and B_v are binary vectors where the count of these a binary vector is based on the number of dimensions on which it has non-zero value [Manning and Schütze, 1999]. The following measures are used to compare the similarity between the two search results [Zhang, 2008*b*].

Matching measure

This measure aims to provide results if the ontology-based search method covers all the results from text-based method. It is similar to the overlap measurement

explained next, but it puts an emphasis on the count of matching results that are retrieved in B specifically.

$$Match(R_v, B_v) = \frac{|R_v \cap B_v|}{|B_v|} \tag{8.12}$$

The value of this measure indicates if the ontology-based method produces, at least, the same results as the text-based method. Other measures will indicate the difference in the results. If the matching result is equal to 100%, that means that $B \subseteq R$.

Overlap coefficient similarity measure

The measure aims to measure the inclusion of results. It is similar to the matching measure but with no emphasis on any of the two search results.

$$Overlp(R_v, B_v) = \frac{|R_v \cap B_v|}{min(|R_v|, |B_v|)} \tag{8.13}$$

The value of this measure is between 0 and 1. The only case that the result is 1, is that when the two search results R and B are exactly the same.

Jaccard co-efficient similarity measure

This measure aims to to compare the similarity between the two search result sets R and B. It emphasizes the similarity between the two vectors with respect to the total number of results.

$$Jacc(R_v, B_v) = \frac{|R_v \cap B_v|}{|R_v \cup B_v|} \tag{8.14}$$

The value of the Jaccard coefficient is between 0 and 1, where 1 indicates that the two vectors are the same while 0 indicates the exact opposite.

8.5 Evaluation Results

Several random queries were performed to test and validate the results of the ontology-based query searching method proposed. The queries were conducted by simulating the query searching process where random ontology concepts are selected and sent via the web query interface. The results of these experiments were compared with the results of running the same queries in the text-based matching method. The ontology concept initiated the query in the ontology-based method is expanded into a vector of related classes that build a context around the ontology concept searched. In text-based searching, the ontology concept label is searched in the RDF store, and exact matching results are retrieved. The set of queries involved in this experiment are based on searching one ontology class, and the simulation process did not support queries composed of Boolean expressions. The results of the ontology-based query searching by the set R are compared against the search results of the text-based search denoted by the set B. The results of the comparison were based on the evaluation criteria discussed in the previous section. The evaluation criteria results are detailed in Table 8.3.

In most cases, the results of these experiments show that the size of the search result set R is greater than the size of B. In other words, the size of the results retrieved with the ontology-based query searching techniques is in most cases higher than the results of text-based matching. The results are proved by the values of calculating the matching measure. It calculates if the ontology-based search results cover all the text-based searching results. The value of the matching measure is equal to 1 if $B \subseteq R$, and it is the case for some of the queries (queries no. 4, 7, 8, 23, 24) and all these cases the size of the search results of both query searching method are the same. In other cases, the matching measure is very high despite the significant difference in the size of search results retrieved by both methods, and it is the case for queries (queries no. 5, 21, 25). The justification of the results for such cases is

based on the process of expanding the ontology concept searched into a query vector. For example, when running a query to search for the ontology concept *"Kidney disease"* (query no. 21), the query vector is created based on the relations between the ontology concepts as stored in the RDF store. The query vector created after expanding the query about *"Kidney disease"* was explained in the previous sections

Table 8.3: Similarity measures comparing the search results of R and B

No.	Query	$\|R\|$	$\|B\|$	$Match$	$Overlp$	$Jacc$
1	Abdominal	537	540	0.91	0.92	0.85
2	Adhesion	81	106	0.75	0.99	0.75
3	Antibody	1127	285	0.89	0.89	0.22
4	Appendicitis	68	68	1.00	1.00	1.00
5	Biopsy	425	278	0.96	0.96	0.61
6	Cardiac arrhythmia	174	21	0.58	0.58	0.04
7	Chest pain	67	67	1.00	1.00	1.00
8	Clostridium	26	26	1.00	1.00	1.00
9	Colostomy	13	15	0.87	1.00	0.87
10	Dehydrogenase	85	83	0.96	0.96	0.91
11	Embolism	102	86	0.95	0.95	0.77
12	Fever	204	194	0.91	0.91	0.80
13	Fibrosis	183	182	0.99	0.99	0.97
14	Gastric ulcer	16	16	0.81	0.81	0.68
15	Heart disease	160	68	0.90	0.90	0.37
16	Hemangioma	18	13	0.77	0.77	0.48
17	Hernia	144	149	0.94	0.97	0.92
18	Influenza	66	63	0.89	0.89	0.77
19	Insulin	158	188	0.84	0.99	0.83
20	Insulinoma	4	5	0.80	1.00	0.80
21	Kidney disease	346	115	0.97	0.97	0.32
22	Leukemia	78	63	0.95	0.95	0.74
23	Lymphangioma	10	10	1.00	1.00	1.00
24	Neck pain	6	6	1.00	1.00	1.00
25	Necrosis	817	320	0.94	0.94	0.36
26	Neoplasm	834	156	0.65	0.65	0.11
27	Oxidase	322	144	0.42	0.42	0.15
28	Plasma	375	442	0.84	0.99	0.84
29	Vasculitis	87	53	0.98	0.98	0.59
30	Virus	351	379	0.80	0.87	0.72
Average		229	138	0.88	0.91	0.68

and detailed in Table 8.1. Some ontology concepts can have related concepts that convey the same meaning and thus widen the search process. Some queries have indicated low matching measures, as is the case for queries (queries no. 26, 27), despite having a significant difference in the size of the search results between the two methods were the largest result retrieved by the ontology-based query searching method. These low matching values suggest that some EMOs in the RDF store might be partially enriched with ontology concepts. The process of annotating the EMOs was developed using an external API incorporated into the LEMO system. Possible API requests might fail due to connection issues or failure to complete the request from the API side. The LEMO system can overcome the issue by resending the request to annotate the same text values or perform periodically updates for the RDF store. It is one of the limitations of the LEMO system that will be explored in the future. Despite such query cases, the average of the matching measure (0.88) indicated a positive sign that the proposed ontology-based query searching method outperformed text-based search techniques on exposing more EMOs to the search process, without considering the relevance of the search results. That was the aim of using ontology-based searching by expanding the search query based on ontology concepts relations that convey the context of the concept. The process of matching and ranking the results is responsible for ordering the results according to their relevance to the query initiated. Measuring the relevance of the search results is beyond the scope of the research and will be considered in the future.

The overlap measure is similar to the matching measure but without putting an emphasis on one of the search results sets. The values of the overlap measure, defined in the previous section, are equal to the matching measure when $|R| > |B|$. In Table 8.3 detailing the results of the query experiments, 8 out of 30 queries (queries no. 1, 2, 9, 17, 19, 20, 28, 30) had the opposite case where $|B| > |R|$. The reason for such cases was explained previously indicating that some EMOs might not have been annotated properly. Thus, the overlap coefficient calculated in

such cases is high, and that indicates that the two search results sets complement each other. The average of the overlap coefficient is high (0.91), and that shows that combining the two methods for query searching is a solution to consider in the future. Moreover, the text-based searching can be applied to searching new elements introduced in the LEMO metadata schema such as the ontology concepts' synonyms that are retrieved and stored in the Term resources description in the RDF store as part of the enrichment process.

As for the Jaccard similarity coefficient, the values are high when the difference between the sizes of the two search results, $|R|$ and $|B|$, is not high. Such values are logical when calculating binary vector similarity. This measure can be used to investigate when one of the methods outperforms the other in retrieving more search results. Having a close number of results retrieved by both methods dictates high Jaccard values, and its close to 1 if the search results are exactly similar as in the case of (query no. 13). The Jaccard coefficient value can be low despite having high overlap values between the two search results compared, as in the case of queries (queries no. 21, 26). The difference in the size of the two search results in these two cases affects the Jaccard coefficient measurement as the number of zero entries increases in one of the vector entries. The average value of the Jaccard coefficient is not high (0.68), and that indicates that one method outperformed the other method in the size of the retrieved search results. The high average of the matching measure indicates that the ontology-based query searching method was able to enhance the discoverability of EMOs described in LEMO metadata schema and stored in the RDF store.

8.6 Summary

An important technique for information retrieval is query searching, and it plays a vital role in any information retrieval system. This chapter has proposed an

ontology-based query searching technique that exploits the Linked Data format applied for building the LEMO dataset. The proposed technique has utilised the LEMO metadata schema elements concerned with describing the hierarchical relations between the ontology concepts used to enrich the LEMO dataset. The focus of this query searching method was to enhance the discoverability of the EMOs by extending the search query to include related concepts to build a contextualised query. This chapter has explained the ontology-based query searching method proposed. It consists of two processes that start with expanding the query initiated to include related concepts that widen the search in the LEMO dataset. The second process explained the matching of EMOs stored in the RDF store with the query initiated. This chapter has presented the results of the proposed ontology-based query searching method compared to the results of the text-based query searching. The aim of the ontology-based query searching was to enhance the discoverability of the EMOs, and that was supported by the results explained in this chapter. The overall values for the overlap between the search results of the two experiments have shown a larger number of EMOs retrieved in ontology-based query searching method in most cases. However, the proposed ontology-based query searching method has some limitations that were explained in this chapter, and will be dealt with in the future research.

In conclusion, this chapter has addressed the research objective **O10**: "Develop an ontology-based query searching algorithm for testing and comparing of query searching results between ontology-based and text-based searching methods in the LEMO dataset". The process of addressing this research objective has completed the answer of the research question **R4**: "How can the Linked Data practices be utilised in the process of accessing and querying the dataset of integrated EMOs called the LEMO dataset? Moreover, how can the linkages between content retrieved from the LEMO dataset be evaluated?". The browsing and query searching methods proposed in chapters 7 and chapter 8 have exploited the elements of the

LEMO metadata schema that was implemented in Linked Data. The RDF store was enriched with concepts from the SNOMED CT ontology and that was utilised for enhancing the search and discoverability of EMOs stored in the LEMO dataset. The retrieval methods have exploited the SNOMED CT hierarchical relations to navigate the LEMO dataset and query its content. The results of ontology-based query searching method developed have proved that it can be beneficial to enhance the search results of incorporated with text-based retrieval methods in the future. More details about the findings of this research and the future research directions are presented in the final chapter of this thesis.

Chapter 9

Conclusions and Future Work

9.1 Conclusions

The open educational resources provided on the internet have changed teaching and learning in general. The emergence of Web 2.0 technologies has also changed what can be considered an educational content. Learner can acquire knowledge nowadays not only from books and articles but also from learning object that is publicly available on the Internet such as videos, blogs, simulations, case studies, and even pictures. The wide range of open data available on the web has made searching for content, that can be used to learn a particular topic, a time-consuming task. The work presented in this thesis have proposed and applied a practical solution to the problem identified in this research in the field of medical education as a proof of concept. The solution has explored techniques and methods for harvesting, enriching, and linking the educational content of different types that is published on the web. In this thesis we have proposed a metadata schema that accommodates describing articles, videos, or blogs published on the web and we introduced new elements that accommodate describing possible enrichments added to the metadata. Furthermore, we have proposed a novel system that harvests and maps educational content from the web into a unified metadata schema, and then, it enriches that

metadata with semantics which will result in having one coherent dataset of educational objects harvested from distributed web data sources. This research has exploited the latest technique for publishing data on the web that is Linked Data (chapter 2, section 2.4.5) which has paved the way towards the emergence of the "Web of Data". In this thesis, the metadata schema proposed has been implemented in Linked Data where new features have been introduced to enable a richer description of educational objects. Moreover, the system proposed in this thesis has implemented methods that create, update, and store Linked Data in an RDF store that manages all the harvested educational objects. The final RDF store resulted from this research was composed of Educational Medical Objects (EMOs) enriched with ontology concepts that enable building linkages between these objects. The RDF store content has been evaluated via techniques developed for accessing and retrieving EMOs from the RDF store. These techniques have exploited the new features proposed in the metadata schema for describing ontology concepts used for annotating and enriching the metadata to enhance browsing and querying the RDF store. The evaluation methods have been useful for validating the new techniques and methods proposed in this work for aggregating and integrating distributed educational objects from the web. One limitation in the evaluation of this work is not involving expert users. Further experiments can be conducted with smaller clusters of datasets where linkages between the EMOs can be evaluated with expert users to validate these linkages.

The chapter aims to review the research conducted in this thesis, discusses the research objectives and contributions achieved, and suggests the possible enhancements that can be undertaken in the future. This research has focused on the medical education domain as a case study of this work that is the medical education. The overall research presented in this thesis have investigated the practices of Linked Data and how they can be exploited to answer the main research question:

R0: How can Linked Data be used to support the acquisition and integration of

EMOs from distributed web data sources? The EMOs represent any piece of information that can be used for learning in the field of medical education such as articles, videos, and blogs.

The process of answering this research question has provided techniques and methods that utilise Linked Data and biomedical ontologies to aggregate and integrate the EMOs from diverse web data sources into one coherent linked dataset. The process of achieving the research objectives and answering the research questions has been reflected on at the end of each chapter in this thesis. Next, we provide a discussion of the research contributions and suggested work to be conducted in the future.

9.1.1 Overview

The main findings of this research are discussed in the rest of this chapter. These findings have been explained throughout the chapters of this thesis. As a recap of the work presented in this thesis, the following list summarises the steps of how this research was conducted as specified in the research design at the beginning of this thesis (Chapter 1, section 1.3).

- **Explore:** the first step in any research project starts by exploring the domain of interest. Chapter 2 presented the necessary background knowledge needed to understand this work, and Chapter 3 presented a domain study that uncovered the problems and frustrations faced by the web users who are concerned with medical education. The gaps discovered in the literature and the recommendations deduced from the domain studied drew the plan for developing a solution to the problem investigated

- **Develop:** the solution proposed in this research included the LEMO metadata schema and the LEMO system that were presented in Chapter 4 and

Chapter 5 respectively. The decisions taken to design and develop the metadata schema and the system were based on the input data collected in the exploratory research. The two chapters have explained the process of designing, developing, and testing each component to run a larger experiment for validating the solution and its effectiveness in solving the problem investigated.

- **Evaluate:** the solution developed in this research is intended to present a coherent dataset of EMOs collected from distributed web data sources. Chapter 6, chapter 7, and chapter 8 presented the results of running the system with data collected from the web, and the experiments conducted to evaluate these results. The results were evaluated using retrieval techniques developed to exploit the Linked Data features for linking data. The evaluation is divided into two stages. Firstly, Chapter 6 evaluated if the solution proposed has succeeded in aggregating and integrating data from web data sources. The integration in this chapter is intended to mean creating linkages between the EMOs. The integration of EMOs together into one linked dataset is validated in chapter 7 and chapter 8 by accessing the linked dataset via ontology-based techniques for browsing and querying its the dataset.

After this recap of the thesis chapters, the research contributions are discussed in the following sections.

9.1.2 The LEMO Metadata Schema

To aggregate EMOs of different types such as videos, blogs, and articles, a simple metadata schema is proposed to accommodate the needs of different types of EMOs that is interoperable with the original metadata of EMOs. The metadata schema is developed as a Dublin Core Application Profile (DCAP) to ensure better interoperability of data. The DC metadata schema has a simple structure. Thus, allowing its elements to be extended to have new features that are used for enriching the

description of metadata elements. In our proposed LEMO metadata schema, the focus was on enhancing the discoverability of the EMOs by adding elements that enrich the description of the textual parts of the metadata such as the Title and the Description metadata elements. The DC metadata schema was adopted because of its simple structure that can be easily extended and because it is interoperable with the possible metadata harvested by the LEMO system. The data sources involved in the LEMO system were web 2.0 websites that describe its content using a simple metadata schema which applies some of the DC metadata elements. The proposed LEMO metadata schema has proved its ability to accommodate describing videos harvested from *YouTube* channels and blogs published in blogging platforms. The LEMO system harvests articles hosted in *PubMed* library and that library uses the NLM metadata schema that is based on the DC metadata schema. Hence, the proposed LEMO metadata schema was able to map to the metadata elements of the data harvested using the LEMO system via XSLT techniques. The proposed LEMO metadata schema was used to describe more than 10,000 EMOs of varying types. The new features introduced in the LEMO metadata schema were represented as either new elements extending the original DC elements or new resources used to describe the enrichments added to the original metadata elements. The LEMO metadata schema is implemented in RDF/XML format, and that enabled having a flexible metadata schema that can be enriched with external ontologies to enhance the description of EMOs. For example, the *title annotation* attribute is a feature introduced to extend the Title metadata element. Another example is the *Term resource* that can describe an ontology concept that annotates the title. The LEMO metadata schema was proposed with the aim of introducing a Linked Data infrastructure that enables the automatic integration of EMOs into one coherent dataset. The refinements introduced in the LEMO schema were elements that are used for describing semantics annotations discovered in the textual elements of EMOs that are the title and description. The semantic annotations are described

in LEMO metadata schema as RDF resources that are related to another RDF resources describing the title or the description of the EMO. Providing such flexibility for describing EMOs has proved its effectiveness in creating linkages between the EMOs based on the hierarchical relations between the ontology concepts annotating it. The main features presented in this research contribution can be summarised as follows.

- The LEMO metadata have new attributes that extend the DC metadata elements defined by the prefix `(lemo:)` and new attributes that are used to describe new RDF resources introduced in the LEMO metadata schema.
- The LEMO metadata have new elements used for describing text annotations and linking it with external ontologies that are annotated in its metadata.
- The LEMO metadata have new elements that store the linkages between the ontology concepts used for annotating the metadata.

9.1.3 The LEMO System

This research introduced the LEMO system which consisted of several processes that were designed to exploit the features of the LEMO metadata schema and harvest the EMOs from diverse web data sources to have one linked dataset. The LEMO system introduced two harvesting endpoints that were responsible for harvesting EMOs using RSS feeds and OAI-PMH protocol from the web. Research efforts aiming at integrating educational data focused on integrating static datasets collected from online libraries in the same domain [Mitsopoulou et al., 2011] or integrating data published on the same platform such as *YouTube* [Fernandez et al., 2011]. The harvesting endpoints introduced in the LEMO system are similar as the RSS feeds reader technique. They collect data from diverse web data sources when given the needed URL. Any website that provides its content as RSS feeds can have its content collected by the LEMO system. Also, any repository that is set up with OAI-PMH

protocol can have its content harvested using the LEMO system.

The second process introduced in the LEMO system was responsible for mapping all the XML metadata files collected into the LEMO metadata schema. The importance of this process was converting all the XML metadata files to RDF/XML format that are stored in one repository. This process has described the EMOs metadata using Linked Data format and has prepared the infrastructure for adding enrichments that metadata. In the LEMO system, the enrichment process exploited the Linked Data for annotating the EMOs with biomedical ontology concepts that added semantics to the text of EMOs. In the experiments conducted to test this process, the results confirmed that automatic linkages can be generated between the EMOs based on the ontology concepts enriching their metadata. The linkages can be created and explored based on the relations between the ontology concepts. The main features presented in this research contribution can be summarised as follows.

- The LEMO system has two harvesting endpoints that can collect data from the web given specific URLs of the sources. Then, it maps the metadata collected into one unified schema that is the LEMO metadata schema represented in RDF/XML format.

- The LEMO system has the ability to annotate the textual elements of the metadata with concepts from biomedical ontologies using BioPortal annotator API. The system updates the metadata of EMOs to enrich its description with external data.

- The LEMO system stores the enriched metadata in an RDF store along with the ontology concepts description that illustrates their relations. The LEMO system prepares a Linked Data infrastructure for accessing and retrieving data from the RDF store.

9.1.4 The LEMO Dataset

One of the main contributions of this research is the final results of running the LEMO system. The result of testing the LEMO system is a repository of enriched EMOs metadata described as Linked Data stored in an RDF store. The dataset consists of metadata of EMOs and their URLs that point to the original EMOs where they are published. The LEMO dataset is exposed as Linked Data and connected to the SNOMED CT ontology that is used to annotate the LEMO dataset. The work presented in this research confirms that the dataset can be extended easily to include more web data sources using the LEMO system harvesting endpoints. Experiments in chapter 5 present results for a smaller dataset used for testing the LEMO System. In chapter 6, the dataset is extended to include more EMOs resulting in having a larger dataset of EMOs of different types and various topics. It was significant in this research that the final LEMO dataset contains a large number of EMOs to be accessed for evaluation. The main features presented in this research contribution can be summarised as follows.

- The LEMO dataset presents a linked dataset of EMOs collected from distributed web data sources that include the *PubMed Library*, *YouTube* channels, and blog articles.
- The LEMO dataset have been enriched with terms annotating its textual description using ontology concepts from the SNOMED CT ontology. The enriched dataset is stored in an RDF store.
- Links can be generated within the LEMO dataset components based on the hierarchical relations of the ontology concepts added to the RDF store.

9.1.5 Ontology-based Retrieval

The evaluation of the LEMO dataset was conducted to explore its content and thus present its linkages and coherence. All the methods developed for accessing the

LEMO dataset exploited the Linked Data features. After storing the LEMO dataset in an RDF store, this research proposed techniques for browsing and querying the RDF store that utilise the features introduced in the LEMO metadata schema. The experiments conducted in this research presented results of accessing the RDF store in addition to comparisons of results that confirms the efficiency of using the LEMO metadata schema and the LEMO system for describing and enriching the EMOs. The browsing technique proposed for exploring the RDF store was based on browsing the ontology concepts used for enriching its content. This technique has been used to evaluate the categorisation of EMOs into ontology concepts. The experiments conducted for clustering of the browsing results have indicated the efficiency of the linkages built between the EMOs. On the other hand, the ontology-based query searching technique developed in this research has also exploited the LEMO metadata schema to improve the process of searching. It utilised the hierarchical relations between the ontology concepts to expand the search query and enhance the discoverability of EMOs. The results of applying this technique were compared with text-based searching. The results have shown significant improvements when searching for some terms, maintained the same results in some cases, and in other cases, it was short in results compared to the text-based search. On average, the results have shown almost 91% of overlap between the results of searching using the two techniques. Furthermore, the results have shown that the two searching techniques can be used to complement each other and enhance the discoverability of EMOs in the RDF store. The main features presented in this research contribution can be summarised as follows.

- Ontology-based browsing technique that evaluated the linkages between the EMOs in the RDF store and validated its coherence via clustering the similar EMOs when retrieved.

- Ontology-based query searching technique that enabled wider retrieval of EMOs

when searching for a concept and by doing that enhances the discoverability of EMOs.

9.2 Suggestions for Future Work

The findings presented in this research present many potential interesting opportunities for future work concerning the delivery of EMOs to web users and improving the techniques developed for building the RDF store. The work presented in this thesis proposed a novel system that builds on top of a metadata schema that introduced new attributes and features for organising diverse types of educational content in the field of medical education. Based on the findings discovered in this research, we suggest the following directions for the future research.

1. The LEMO metadata schema describes the people involved in developing the educational objects, the usage rights for it, along with other elements that describe the content of the object such as its type, title, and description. This schema was developed in RDF/XML format that enables the use of Linked Data features for enriching the metadata. In this research, the use of biomedical ontologies such as SNOMED CT and MeSH have been implemented to enrich the title and the description of objects with semantics that enables its linking with each other and enhances its discovery when searching. The metadata can be improved using other ontologies such as FOAF ontology for describing the creators, publishers, or contributors of the educational resources. Furthermore, usage rights metadata element can be described with Common Creative vocabulary to declare the licensing of the resources published. The process of using such ontologies will link the dataset to external sources that expose it to be linked by other datasets.

2. The process of enriching the metadata can be improved by enabling the use of multiple biomedical ontologies for annotating the textual description provided

in the metadata of the educational resources. The LEMO metadata schema is ready for handling such process. It can describe annotations added to the text using different ontologies. Utilising multiple ontologies can provide depth for the dataset and provide different layers for browsing the dataset based on the concepts of each ontology. Furthermore, the system can incorporate techniques for ontology alignment in order to match the corresponding concepts in two ontologies and eliminate any redundancy in the annotations. This will help in having a richer annotations as some ontologies compliment each other.

3. The LEMO RDF store content can be evaluated with users accessing the store. Several experiments can be conducted with expert users for evaluating the quality of the annotations discovered for the EMOs. Further evaluation of the retrieval methods can be conducted by collecting users feedback when browsing and querying the RDF store.

4. The techniques developed for retrieving EMOs from the RDF store can be enhanced to exploit the hierarchical relations between the concepts of an ontology. New features that can enable the user to specify the scope of the retrieval process can enhance the discoverability of EMOs. The ontology-based retrieval methods are based on the hierarchical relations between the ontology concepts. Hence, the retrieval process can be controlled by expanding the matching of EMOs to a circle of related concepts instead of one concept when the user interacts with the system. Furthermore, the user can be introduced with an option to select the number of levels they want to expand when interacting with the system.

5. The web interface for accessing the RDF store can be implemented as a website with other features that can be used by the web users to grow the dataset. For example, allowing the users to add RSS feeds URLs to read in the future. In such cases, the system must have administrative team for authorising the web

data sources entered. Furthermore, rules and regulations for harvesting and enriching the data must be incorporated to obtain the best performance of the system without affecting the users access. New open web data sources can be harvested and added to the LEMO dataset by the administrative team after providing the required XSLT files for mapping. Then, the content of the RDF store can be updated periodically for enriching new EMOs harvested by users. For example, the harvesting and enriching of recent new web data sources can be performed monthly and in off-peak hours.

6. Browsing and query searching methods introduced in this research can be improved and fully implemented as part of that website. The website can be enhanced to include the same features introduced in the LEMO system such as the harvesting endpoints. Users can enter URLs of RSS feeds they want to syndicate and collect educational resources that can be enriched and added to the collection of EMOs stored in the RDF store. Also, the system can be developed to have adaptive content based on users interests. It can be easily developed to view only specific RSS feeds or EMOs annotated with particular subjects. A bold step for this work in the future is to develop a service on this website that provides an adaptive delivery for EMOs to its users via social networks. Twitter is considered as a possible delivery tool to be incorporated in the future of developing the website. The adaptive delivery of EMOs via Twitter would be a fascinating and different research area to explore.

7. The same techniques used in this research can be applied to harvest and enrich educational materials from other disciplines. With the use of proper ontologies specialised in the new discipline and with developing the needed API for annotating the metadata, the same techniques can be used to build a dataset for another discipline.

References

Abel, F., Celik, I., Hauff, C., Hollink, L. and Houben, G. [2011], U-Sem: Semantic Enrichment, User Modeling and Mining of Usage Data on the Social Web, *in* 'USEWOD2011, proceedings of 1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011), workshop at the 20th International World Wide Web Conference (WWW2011)'.

Abelson, H., Adida, B., Linksvayer, M. and Yergler, N. [2008], ccREL: The creative commons rights expression language, Technical report, Technical report, Creative Commons, 2008. http://wiki. creativecommons. org/Image: Ccrel-1.0. pdf.

Adida, B. [2008], 'hgrddl: Bridging microformats and {RDFa}', *Web Semantics: Science, Services and Agents on the World Wide Web* **6**(1), 54 – 60.

Adida, B., Birbeck, M., McCarron, S. and Pemberton, S. [2008], 'RDFa in XHTML: Syntax and processing', `https://www.w3.org/MarkUp/2008/ED-rdfa-syntax-20081004/rdfa-syntax.pdf`. Online. Accessed on April 25, 2014.

Alonso-Roris, V. M., Míguez-Pérez, R., Santos-Gago, J. M. and Álvarez-Sabucedo, L. [2012], A semantic enrichment experience in the early childhood context, *in* 'Frontiers in Education Conference (FIE), 2012', IEEE, pp. 1–6.

Bamidis, P. D., Kaldoudi, E. and Pattichis, C. [2009], mEducator: a best practice network for repurposing and sharing medical educational multi-type content, *in* 'Leveraging Knowledge for Innovation in Collaborative Networks', Springer, pp. 769–776.

Barker, P. and Campbell, L. M. [2014], 'What is schema. org?', `http://publications.cetis.org.uk/wp-content/uploads/2014/06/schemaBriefing.pdf`. Online. Accessed on December 20, 2015.

Benesty, J., Chen, J., Huang, Y. and Cohen, I. [2009], Pearson correlation coefficient, *in* 'Noise reduction in speech processing', Springer, pp. 1–4.

Bennett, S., Bishop, A., Dalgarno, B., Waycott, J. and Kennedy, G. [2012], 'Implementing web 2.0 technologies in higher education: A collective case study', *Computers & Education* **59**(2), 524–534.

Benslimane, D., Dustdar, S. and Sheth, A. [2008], 'Services mashups: The new generation of web applications', *IEEE Internet Computing* **12**(5), 13.

Bergman, M. K. [2001], 'White paper: the deep web: surfacing hidden value', *Journal of electronic publishing* **7**(1).

Berners-Lee, T. [2006], 'Design issues: Linked data', `https://www.w3.org/DesignIssues/LinkedData.html`. Online. Accessed on April 23, 2014.

Berners-Lee, T., Fielding, R. and Masinter, L. [2005], Uniform resource identifier (uri): Generic syntax, Technical report. Online. Accessed on April 24, 2014.

Bertini, M., Devedi, V., Gaevi, D. and Torniai, C. [2011], 'Guest editorial: Semantic technologies for multimedia-enhanced learning environments', *Interactive Learning Environments* **19**(1), 1–4.

Bizer, C., Heath, T. and Berners-Lee, T. [2009], 'Linked data-the story so far', *International journal on semantic web and information systems* **5**(3), 1–22.

Bizer, C., Heath, T., Idehen, K. and Berners-Lee, T. [2008], Linked data on the web (ldow2008), *in* 'Proceedings of the 17th international conference on World Wide Web', ACM, pp. 1265–1266.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S. [2009], 'DBpedia-A crystallization point for the Web of Data', *Web Semantics: science, services and agents on the world wide web* **7**(3), 154–165.

Bodenreider, O. [2004], 'The unified medical language system (UMLS): integrating biomedical terminology', *Nucleic acids research* **32**(1), 267–270.

Bratsas, C., Chrysou, D. E., Eftychiadou, E., Kontokostas, D., Bamidis, P. and Antoniou, I. [2012], Semantic web game based learning: An i18n approach with greek dbpedia, *in* 'Proceedings of the 2nd International Workshop on Learning and Education with the Web of Data (LiLe-2012 at WWW-2012), Lyon, France'.

Brickley, D. and Guha, R. V. [2014], 'Rdf schema 1.1', *W3C Recommendation* . Online. Accessed on July 20, 2015.

Brickley, D. and Miller, L. [2012], 'FOAF vocabulary specification 0.98', *Namespace document* **9**.

Brown, S. A. [2012], 'Seeing web 2.0 in context: A study of academic perceptions', *The Internet and Higher Education* **15**(1), 50–57.

Büttcher, S., Clarke, C. L. and Cormack, G. V. [2010], *Information retrieval: Implementing and evaluating search engines*, Mit Press.

Candler, C. S., Uijtdehaage, S. H. and Dennis, S. E. [2003], 'Introducing HEAL: The health education assets library', *Academic medicine* **78**(3), 249–253.

Carmichael, P. and Jordan, K. [2012], 'Semantic web technologies for education–time for a turn to practice?', *Technology, Pedagogy and Education* **21**(2), 153–169.

Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P. and Quarteroni, S. [2013*a*], Classification and clustering, *in* 'Web Information Retrieval', Data-Centric Systems and Applications, Springer Berlin Heidelberg, pp. 39–56.

Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P. and Quarteroni, S. [2013*b*], Publishing data on the web, *in* 'Web Information Retrieval', Data-Centric Systems and Applications, Springer Berlin Heidelberg, pp. 137–159.

Chang, K. C.-C. and Cho, J. [2006], Accessing the web: from search to integration, *in* 'Proceedings of the 2006 ACM SIGMOD international conference on Management of data', ACM, pp. 804–805.

Couldry, N. [2012], *Media, society, world: Social theory and digital media practice*, Polity.

Coyle, K. [2005], 'Understanding metadata and its purpose', *The Journal of Academic Librarianship* **31**(2), 160–163.

Coyle, K. and Baker, T. [2009], 'Guidelines for dublin core application profiles', `http://dublincore.org/documents/profile-guidelines/`. Online. Accessed on November 20, 2015.

CWA, C. W. A. [2006], Guidelines and support for building application profiles in e-learning, Technical report.

Cyganiak, R. and Jentzsch, A. [2011], 'Linking open data cloud diagram', `http://lod-cloud.net/`. Online. Accessed on June 5, 2015.

Cyganiak, R., Wood, D. and Lanthaler, M. [2014], 'RDF 1.1 concepts and abstract syntax', *W3C Recommendation* . Online. Accessed on July 20, 2015.

dÁquin, M. [2012], Putting linked data to use in a large higher-education organisation, *in* 'Proceedings of the Interacting with Linked Data (ILD) workshop at Extended Semantic Web Conference (ESWC)'.

David, C., Kohlhase, M., Lange, C., Rabe, F., Zhiltsov, N. and Zholudev, V. [2010], Publishing math lecture notes as linked data, *in* 'The Semantic Web: Research and Applications', Springer, pp. 370–375.

de Carvalho Moura, A. M., Campos, M. L. M. and Barreto, C. M. [1998], 'A survey on metadata for describing and retrieving internet resources', *World Wide Web* **1**(4), 221–240.

Dennis, S. E., Dippie, S. R., Candler, C. S., McIntyre, S. A. and Uijtdehaage, S. [2004], An indexing standard for sharing health education multimedia resources: The health education assets library (heal) metadata schema, *in* 'System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on', IEEE, pp. 138–147.

Devedžic, V. [2006], *Semantic web and education*, Vol. 12, Springer Science & Business Media.

Diamantopoulos, N., Sgouropoulou, C., Kastrantas, K. and Manouselis, N. [2011], Developing a metadata application profile for sharing agricultural scientific and scholarly research resources, *in* 'Metadata and Semantic Research', Springer, pp. 453–466.

Dietze, S., Sanchez-Alonso, S., Ebner, H., Qing Yu, H., Giordano, D., Marenzi, I. and Pereira Nunes, B. [2013], 'Interlinking educational resources and the web of data: A survey of challenges and approaches', *Emerald Program: electronic Library and Information Systems* **47**(1), 60–91.

Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N. and Taibi, D. [2012], Linked education: interlinking educational resources and the web of data, *in* 'Proceedings of the 27th annual ACM symposium on applied computing', ACM, pp. 366–371.

Duval, E., Hodgins, W., Sutton, S. and Weibel, S. L. [2002], 'Metadata principles and practicalities', *D-lib Magazine* **8**(4), 16.

El-Sherbini, M. and Klim, G. [2004], 'Metadata and cataloging practices', *The Electronic Library* **22**(3), 238–248.

Elevitch, F. R. [2005], 'SNOMED CT: electronic health record enhances anesthesia patient safety', *AANA journal* **73**(5), 361.

Fernandez, M., dÁquin, M. and Motta, E. [2011], Linking data across universities: an integrated video lectures dataset, *in* 'The Semantic Web–ISWC 2011', Springer, pp. 49–64.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and Berners-Lee, T. [1999], 'Hypertext transfer protocol–HTTP/1.1', `http://www.hjp.at/doc/rfc/rfc2616.html`. Online. Accessed on January 21, 2016.

Fleiszer, D. M., Posel, N. H. and Steacy, S. P. [2004], 'New directions in medical e-curricula and the use of digital repositories', *Academic Medicine* **79**(3), 229–235.

Franks, P. and Kunde, N. [2006], 'Why metadata matters', *Information Management* **40**(5), 55.

Friesen, N. [2001], 'What are Educational Objects?', *Interactive Learning Environments* **9**(3), 219–230.

Ghanem, T. M. and Aref, W. G. [2004], 'Databases deepen the web', *Computer* **37**(1), 116–117.

Gorman, M. [2003], 'Cataloguing in an electronic age', *Cataloging & Classification Quarterly* **36**(3-4), 5–17.

Granichin, O., Volkovich, Z. and Toledano-Kitai, D. [2015], Cluster validation, *in* 'Randomized Algorithms in Automatic Control and Data Mining', Vol. 67 of *Intelligent Systems Reference Library*, Springer Berlin Heidelberg, pp. 163–228.

Guenther, R. and Radebaugh, J. [2004], 'Understanding metadata', *National Information Standard Organization (NISO) Press, Bethesda, USA* . Online. Accessed on November 26, 2015.

Haslhofer, B. and Klas, W. [2010], 'A survey of techniques for achieving metadata interoperability', *ACM Computing Surveys (CSUR)* **42**(2), 7.

Heath, T. [2008], 'How will we interact with the web of data?', *Internet Computing, IEEE* **12**(5), 88–91.

Heath, T. and Bizer, C. [2011], 'Linked data: Evolving the web into a global data space', *Synthesis lectures on the semantic web: theory and technology* **1**(1), 1–136.

Heery, R. and Patel, M. [2000], 'Application profiles: mixing and matching metadata schemas', *Ariadne* **25**.

Hendler, J. [2008], 'Web 3.0: Chicken farms on the semantic web', *Computer* **41**(1), 106–108.

Hendrix, M., Protopsaltis, A., Dunwell, I., de Freitas, S., Petridis, P., Arnab, S., Dovrolis, N., Kaldoudi, E., Taibi, D., Dietze, S. et al. [2012], Technical evaluation of the meducator 3.0 linked data-based environment for sharing medical educational resources, *in* '2nd International Workshop on Learning and Education with the Web of Data, Lyon, France', Vol. 4.

Hodgson, C. [2008], 'Building a metadata schema where to start'. Online. Accessed on November 25, 2015.

Hoehndorf, R., Dumontier, M. and Gkoutos, G. V. [2012], 'Evaluation of research in biomedical ontologies', *Briefings in bioinformatics* .

Horrocks, I. [2008], 'Ontologies and the semantic web', *Communications of the ACM* **51**(12), 58–67.

Hyland, B., Atemezing, G., Pendleton, M. and Srivastava, B. [2014], 'Linked data glossary', *W3C working group note, W3C* . Online. Accessed on December 13, 2015.

IEEE LTSC [2002], 'Ieee standard for learning object metadata', *IEEE Std 1484.12.1-2002* pp. i–32.

Iiyoshi, T. and Kumar, M. V. [2008], *Opening up education: The collective advancement of education through open technology, open content, and open knowledge*, Mit Press.

IMS Global Learning Consortium [2006], 'IMS Meta-data Best Practice Guide for IEEE 1484.12. 1-2002 Standard for Learning Object Metadata, Version 1.3 Final Specification', `https://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html`. Online. Accessed on February 20, 2014.

Isaac, Y., Bourda, Y. and Grandbastien, M. [2012], Semunit-french unt and linked data, *in* 'LiLe-2012 at WWW-2012', Vol. 840, CEUR workshop proceedings, p. 6.

ISO/TC [2009], 'Information and documentation: The dublin core metadata element set', `http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52142`. Online. Accessed on November 13, 2014.

Jeremić, Z., Jovanović, J. and Gašević, D. [2013], 'Personal learning environments on the social semantic web', *Semantic Web* **4**(1), 23–51.

Jonquet, C., LePendu, P., Falconer, S., Coulet, A., Noy, N. F., Musen, M. A. and Shah, N. H. [2011], 'NCBO Resource Index: Ontology-based search and mining of biomedical resources', *Web Semantics: Science, Services and Agents on the World Wide Web* **9**(3), 316–324.

Kaliyaperumal, K. [2004], 'Guideline for Conducting a Knowledge , Attitude and Practice (KAP) Study', *Community Ophthalmology* **4**(1), 7–9.

Khare, R. [2006], 'Microformats: The next (small) thing on the semantic web?', *Internet Computing, IEEE* **10**(1), 68–75.

Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I. and Metakides, G. [2012], 'Internationalization of linked data: The case of the greek {DBpedia} edition', *Web Semantics: Science, Services and Agents on the World Wide Web* **15**, 51 – 61.

Kunze, J. A. and Baker, T. [2007], 'The dublin core metadata element set', `http://tools.ietf.org/html/rfc5013.html`. Online. Accessed on November 15, 2014.

Lagoze, C. and Van de Sompel, H. [2003], 'The making of the open archives initiative protocol for metadata harvesting', *Library hi tech* **21**(2), 118–128.

Lama, M., Vidal, J. C., Otero-García, E., Bugarín, A. and Barro, S. [2012], 'Semantic linking of learning object repositories to dbpedia', *Journal of Educational Technology & Society* **15**(4), 47–61.

Lassila, O. [1998], 'Web metadata: A matter of semantics', *IEEE Internet Computing* (4), 30–37.

Launiala, A. [2009], 'How much can a KAP survey tell us about people's knowledge, attitudes and practices? Some observations from medical anthropology research on malaria in pregnancy in Malawi', *Anthropology Matters* **11**(1).

Learning Technology Standards Committee [2002], 'IEEE Standard for learning object metadata', *IEEE Standard* **1484**(1).

Lee, D., de Keizer, N., Lau, F. and Cornet, R. [2014], 'Literature review of SNOMED CT use', *JAMIA Journal of the American Medical Informatics Association* **21**(E1), 11–19.

Lieberam-Schmidt, S. [2010], Web structure, *in* 'Analyzing and Influencing Search Engine Results', Gabler, pp. 49–103.

Lipscomb, C. E. [2000], 'Medical subject headings (MeSH)', *Bulletin of the Medical Library Association* **88**(3), 265.

Liu, B. [2007], *Web data mining: exploring hyperlinks, contents, and usage data*, Springer Science & Business Media.

Liu, B. [2011], Information retrieval and web search, *in* 'Web Data Mining', Springer, pp. 211–268.

Manning, C. D., Raghavan, P., Schütze, H. et al. [2008], *Introduction to information retrieval*, Vol. 1, Cambridge university press Cambridge.

Manning, C. D. and Schütze, H. [1999], *Foundations of statistical natural language processing*, MIT press.

Martin, S., Diaz, G., Sancristobal, E., Gil, R., Castro, M. and Peire, J. [2011], 'New technology trends in education: Seven years of forecasts and convergence', *Computers & Education* **57**(3), 1893–1906.

Médicins du Monde [2011], 'The KAP Survey Model (Knowledge, Attitudes, and Practices)'. Online. Accessed on October 28, 2015.

Mika, P. [2008], Microsearch: An interface for semantic search., *in* 'Semantic Search, International Workshop located at the 5th European Semamntic Web Conference (ESWC 2008)', pp. 79–88.

Miles, A., Matthews, B., Wilson, M. and Brickley, D. [2005], SKOS core: simple knowledge organisation for the web, *in* 'International Conference on Dublin Core and Metadata Applications', p. 3.

Mitsopoulou, E., Taibi, D., Giordano, D., Dietze, S., Yu, H. Q., Bamidis, P., Bratsas, C. and Woodham, L. [2011], 'Connecting medical educational resources to the Linked Data cloud: the mEducator RDF Schema, store and API', *Proceedings of Linked Learning* **22**.

Naeve, A., Lytras, M., Nejdl, W., Balacheff, N. and Hardin, J. [2006], 'Advances of the semantic web for e-learning: expanding learning frontiers', *British Journal of Educational Technology* **37**(3), 321–330.

Neiswender, C. and Montgomery, E. [2009], 'Metadata interoperabilitywhat is it, and why is it important?.', `http://marinemetadata.org/guides/mdataintro/mdatainteroperability`. Online. Accessed on December 25, 2015.

Neiswender, C. and Montgomery, E. [2011], 'Machine readability', `http://marinemetadata.org/guides/mdataintro/machinereadability`. Online. Accessed on December 25, 2015.

Newland, B. and Byles, L. [2014], 'Changing academic teaching with web 2.0 technologies', *Innovations in Education and Teaching International* **51**(3), 315–325.

Nilsson, M., Baker, T. and Johnston, P. [2008], 'Singapore framework for dublin core application profiles. retrieved, april 30, 2011', `http://dublincore.org/documents/singapore-framework/`. Online. Accessed on Noevember 20, 2015.

Nilsson, M., Johnston, P., Naeve, A. and Powell, A. [2007], 'The future of learning object metadata interoperability', *Learning Objects: Standards, Metadata, Repositories, and LCMS* pp. 255–313.

NLM [2004], 'NLM Metadata Schema', `https://www.nlm.nih.gov/tsd/cataloging/metafilenew.html`. Online. Accessed on April 2, 2014.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G. et al. [2009], 'BioPortal: ontologies and integrated data resources at the click of a mouse'.

Piedra, N., Chicaiza, J., López, J., Tovar, E. and Martinez-Bonastre, O. [2012], Combining linked data and mobiles devices to improve access to ocw, *in* 'Global Engineering Education Conference (EDUCON), 2012 IEEE', IEEE, pp. 1–7.

Pirrotta, G. [2010], Linking italian university statistics, *in* 'Proceedings of the 6th International Conference on Semantic Systems', ACM, p. 2.

Popoiu, M. C., Grosseck, G. and Holotescu, C. [2012], 'What do we know about the use of social media in medical education?', *Procedia-Social and Behavioral Sciences* **46**, 2262–2266.

Powell, A., Nilsson, M., Naeve, A., Johnston, P. and Baker, T. [2007], 'DCMI abstract model'.

Prudhommeaux, E., Seaborne, A. et al. [2008], 'SPARQL query language for RDF', `https://www.w3.org/TR/rdf-sparql-query/`. Online. Accessed on December 17, 2015.

Quilitz, B. and Leser, U. [2008], Querying distributed rdf data sources with sparql, *in* 'The Semantic Web: Research and Applications', Vol. 5021, Springer Berlin Heidelberg, pp. 524–538.

Rebai, I., Dela Pssardiere, B. and Labat, J.-M. [2008], A formalism to describe open standards in order to generate application profiles, *in* 'Proceedings of the Internatiotnal Conference on Computer in Education', Citeseer, pp. 491–498.

Richardson, L. and Ruby, S. [2008], *RESTful web services*, O'Reilly Media, Inc.

Robinson, J., Stan, J. and Ribière, M. [2012], Using linked data to reduce learning latency for e-book readers, *in* 'The Semantic Web: ESWC 2011 Workshops', Springer, pp. 28–34.

Rubin, D. L., Shah, N. H. and Noy, N. F. [2008], 'Biomedical ontologies: a functional perspective', *Briefings in bioinformatics* **9**(1), 75–90.

Ruiz-Calleja, A., Vega-Gorgojo, G., Asensio-Pérez, J. I., Bote-Lorenzo, M. L., Gómez-Sánchez, E. and Alario-Hoyos, C. [2012], 'A linked data approach for the discovery of educational ict tools in the web of data', *Computers & Education* **59**(3), 952–962.

Ruiz-Rube, I., Cornejo, C. M. and Dodero, J. M. [2011], 'Accessing learning resources described in semantically enriched weblogs', *International Journal of Metadata, Semantics and Ontologies* **6**(3-4), 175–184.

Sampson, D. [2004], The evolution of educational metadata: From standards to application profiles, *in* 'Proceedings of the IEEE International Conference on Advanced Learning Technologies', IEEE Computer Society, pp. 1072–1073.

Sampson, D. G., Lytras, M. D., Wagner, G. and Diaz, P. [2004], 'Ontologies and the semantic web for e-learning (guest editorial)', *Journal of Educational Technology & Society* **7**(4), 26–28.

Shadbolt, N., Hall, W. and Berners-Lee, T. [2006], 'The semantic web revisited', *Intelligent Systems, IEEE* **21**(3), 96–101.

Sicilia, M.-A., Ebner, H., Sánchez-Alonso, S., Álvarez, F., Abián, A. and García-Barriocanal, E. [2011], 'Navigating learning resources through linked data: a

preliminary report on the re-design of organic. edunet', *Proceedings of Linked Learning* **2011**, 1st.

Siemsens, G. [2005], 'Connectivism: A learning theory for the digital age', *International Journal of Instructional Technology and Distance Learning* . Online. Accessed on November 25, 2015.

Smothers, V. [2004], 'Healthcare Learning Object Metadata. Specifications and description document', `http://www.medbiq.org/sites/default/files/files/HealthcareLOMSpecifications_pointrelease.pdf`. Online. Accessed on January 13, 2014.

Stearns, M. Q., Price, C., Spackman, K. A. and Wang, A. Y. [2001], Snomed clinical terms: overview of the development process and project status., *in* 'Proceedings of the AMIA Symposium', American Medical Informatics Association, p. 662.

Tiropanis, T., Millard, D. and Davis, H. C. [2012], 'Guest editorial: Special section on semantic technologies for learning and teaching support in higher education', *IEEE Transactions on Learning Technologies* (2), 102–103.

Turner, T. [2002], 'What is metadata', *Kaleidoscope* **10**(7), 1.

Van Hage, W. R., De Rijke, M. and Marx, M. [2004], Information retrieval support for ontology construction and use, *in* 'The Semantic Web–ISWC 2004', Springer, pp. 518–533.

Van Harmelen, F. and McGuinness, D. L. [2004], 'Owl web ontology language overview', *World Wide Web Consortium (W3C) Recommendation* .

Vega-Gorgojo, G., Asensio-Pérez, J. I., Gómez-Sánchez, E., Bote-Lorenzo, M. L., Munoz-Cristobal, J. A. and Ruiz-Calleja, A. [2015], 'A review of linked data proposals in the learning domain', *Journal of Universal Computer Science* **21**(2), 326–364.

Waitelonis, J., Sack, H., Hercher, J. and Kramer, Z. [2010], Semantically enabled exploratory video search, *in* 'Proceedings of the 3rd international semantic search workshop', ACM, p. 8.

Wiley, D. [2011], 'Learning objects, content management, and e-learning', *Content management for e-learning* pp. 43–54.

Yee, R. [2008], Learning web services apis through flickr, *in* 'Pro Web 2.0 Mashups: Remixing Data and Web Services', pp. 121–170.

Yoosooka, B. and Wuwongse, V. [2012], 'Linked open data for learning object discovery in adaptive e-learning systems', *International Journal of Knowledge and Learning 3* **8**(3-4), 188–218.

Yu, H. Q., Pedrinaci, C., Dietze, S. and Domingue, J. [2012], 'Using linked data to annotate and search educational video resources for supporting distance learning', *Learning Technologies, IEEE Transactions on* **5**(2), 130–142.

Yu, L. [2011], *A developers guide to the semantic Web*, Springer Science & Business Media.

Zablith, F., Fernandez, M. and Rowe, M. [2015], 'Production and consumption of university linked data', *Interactive Learning Environments* **23**(1), 55–78.

Zaki, M. J. and Wagner, M. J. [2014], *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press.

Zhang, J. [2008*a*], Information retrieval preliminaries, *in* J. Zhang, ed., 'Visualization for Information Retrieval', Vol. 23 of *The Information Retrieval Series*, Springer Berlin Heidelberg, pp. 21–46.

Zhang, J. [2008*b*], *Visualization for information retrieval*, Vol. 23, Springer.